LegalSumAI: Robust and Representative Legal Summarization with Large Language Models

Anonymous submission

Abstract

Legal disputes are complex, high-stakes challenges, yet 56% of US households encounter legal issues annually. The jargon, length, and complexity of legal documents often make them inaccessible to the general public. This work introduces LegalSumAI, a novel framework that generates robust legal summaries from case documents using Large Language Models (LLMs). We investigate whether LLMs can produce accurate, comprehensible summaries of legal cases and opinions without hallucinations. Our two-step pipeline first generates structured CSV fact sheets, capturing case details via the IRAC framework (Issue, Rule, Application, Conclusion) and metadata on the involved parties. These fact sheets are then converted into natural language summaries prompted with Chain of Density (CoD) techniques. We leverage the Multi-LexSum dataset, which provides expert-authored summaries at three different granularity levels (tiny, short, long) to evaluate the generated summaries using ROUGE and BERTScore metrics. Results show high semantic accuracy, with average BERTScore improvements of 157% over Multi-LexSum baselines, demonstrating that our structured reasoning and CoD prompting mitigate LLM hallucinations and improve legal summarization results. This research demonstrates the potential for interpretable LLM pipelines to democratize access to legal knowledge.

Introduction

Legal matters tend to be high stakes, often involving people's fundamental rights, liberties, or significant financial consequences. In criminal proceedings, lives and incarceration may be at stake, while in civil cases, anywhere between tens of thousands to millions of dollars can be on the line. Given these stakes, individuals without legal backgrounds should have access to accurate, comprehensible information about their cases. However, the wording and structure present significant barriers to accessibility.

A typical legal document is approximately 30 pages on average and is written in dense language, making 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, Stygall (2020) 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." This combination of length, complexity, and intertextual dependence makes legal documents difficult for the general public to understand.

Legal tasks are not only high-stakes and complex, but many American households are actively involved in 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 (Rickard 2018). When considering other court systems, it is evident that a majority of American households grapple with legal issues.

Given the scale and impact of these challenges, there is a clear need for effective simplification and summarizing methods. Large Language Models (LLMs) offer a promising approach to address this issue. By leveraging advanced natural language processing techniques, LLMs can analyze, interpret, and summarize vast amounts of legal information. 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 transformer models like BERT, leading to homophobic and transphobic outputs (Felkner et al. 2022). Furthermore, LLMs also often hallucinate, generating output that is highly illogical based on the input context. In high-stakes legal contexts, such distortions or factual inaccuracies can jeopardize trust and fairness of our legal systems. Thus, a current key challenge involves designing robus, interpretable LLM frameworks that minimize hallucinations and bias and maintain factual accuracy and precision. Our work addresses this need through structured and multi-stage reasoning, designed for the rigor required in legal summarization tasks.

Related Work

Benchmarks and Datasets

Recent advances in natural language processing have led to the creation of several domain-specific legal datasets aimed at improving model understanding and reasoning. LeXFiles is comprised of 11 sub-corpora and supports generalization across different legal fields (Chalkidis et al. 2023). Cambridge Law Corpus contains 250,000 UK court cases from the 16th to 21st centuries, designed for NLP and machine learning studies (Östling et al. 2024). CaseHOLD of-

fers multiple choice data derived from US legal documents, supporting baseline performance evaluations (Zheng et al. 2021). MultiLegalPile is a multilingual dataset with more than 680GB of data across 24 languages, promoting cross-lingual legal research (Niklaus et al. 2023).

While these datasets have advanced the field, they often focus on classification tasks or very specific legal text types. This has left gaps in legal summarization research, demonstrating that it is still young and unexplored.

However, Multi-LexSum, a featured paper at NeurIPS, established one of the first large-scale benchmarks for legal summarization across multiple documents (Shen et al. 2022). It contains more than 9,000 expert-developed summaries at three different granularities (tiny, short, and long). Each summary is factually grounded and serves as the gold standard for research in legal summarization.

Summarization Tasks

Shen et al. (2022) evaluated current state-of-the-art summarization frameworks on the Multi-LexSum benchmark. These included BART, PEGASUS, PRIMERA, and two different Longformer-Encoder-Decoder (LED) models applied across the different granularities. While the PRIMERA model improved recall with shorter document context, it did not perform well with longer documents typical of real-world legal cases. In fact, human evaluators scored model outputs an average of 0.43 on a 0-3 quality scale. The authors concluded "existing summarization models struggle to produce the summaries directly from the long source documents (Shen et al. 2022)," emphasizing issues with current legal summarization methods.

To address these issues, we developed an LLM-based framework, which can better handle longer context lengths and complexities typical of legal cases. We explored structured and interpretable reasoning to avoid hallucinations and provide factual grounding and semantic accuracy for legal summarization tasks.

Methods

This work focused on generating robust legal summaries from case documents using LLMs. The primary task was to prompt an LLM to produce accurate, comprehensible summaries of legal cases and opinions for the general public. To achieve this, we investigated several 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 framework?

Introduction to Multi-LexSum

As mentioned in the previous section, the Multi-LexSum dataset, introduced by (Shen et al. 2022), addressed 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 (Shen et al. 2022). 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 (Shen et al. 2022). We utilized the 9,000 Multi-LexSum summaries as our benchmark.

Framework Overview

Here, we propose a two-step pipeline. Instead of directly prompting the LLM for a natural language summary from the source documents, we first generated a CSV "fact sheet" using LLMs to represent the information in a more structured manner. This fact sheet contains categorical information on the Issue, Rule, Application, and Conclusion, parties involved, and other logistical case information for each document. The resulting fact sheet was then parsed by another LLM prompted using CoD to generate a summary of a specified length and purpose.

Fact Sheet Generation

Our approach began with generating "fact sheets" for each legal case document using LLMs. This step involved 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 ensured that the LLM captured essential elements of each case without being overwhelmed by the document's length and complexity.

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

Through a review of existing legal reasoning practices, we identified IRAC as the dominant analytical framework used by lawyers and law students to organize legal arguments and summarize case reasoning. Incorporating this structure as an intermediate step enabled the model to maintain logical coherence and factual consistency with human legal reasoning.

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 created a fact-sheet from each document individually and then combined them into a single comprehensive fact sheet.

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

Generating Natural Language Summaries with Chain of Density

The LLM was then prompted with the fact sheet as contextual input to generate summaries at three granularities ("tiny," "short," and "long"), corresponding to the Multi-LexSum benchmark levels.

For each summary size, the model was prompted four times using Chain-of-Density (CoD) prompting to generate multiple potential summaries. Within each prompt iteration, five summaries were produced. The best summary from each prompt iteration was then selected based on highest cosine similarity scores, ensuring semantic consistency with previous outputs. The best summary from the last Chain-of-Density prompt was selected. This method helped in creating more focused and dense summaries by iteratively refining the prompts, ensuring that each subsequent summary captured more relevant details.

Evaluation Metrics

To evaluate the accuracy of the generated summaries, we computed 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) score, on the other hand, provides a more nuanced assessment of semantic similarity by comparing the embeddings of the generated and reference summaries. High ROUGE and BERT scores indicate that a summary accurately reflects the content of the source documents, while lower scores highlight areas where the summary may need improvement. This evaluation step was critical to ensure reliability and validity of generated summaries.

Justification for Chosen Approach

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

This approach not only addresses the primary research questions 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 aimed to produce summaries that were not only accurate and comprehensible but also accessible and unbiased, thereby enhancing the usability of LLMs in the legal domain.

Experiments

Framework Summary

This framework 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.

We utilized GPT 3.5 turbo, a pre-trained Large Language Model for all of the summarization and reasoning tasks. The model was prompted using the CoD technique to generate natural language summaries. For each summary size, we prompted the model four times, generating five summaries per prompt. The best summary from each prompt was selected based on cosine similarity, and the final summary was selected from the last CoD prompt.

Performance Metrics

The performance of the generated summaries was evaluated using ROUGE and BERTScore metrics. 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.

	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 1: 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

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

Our results had high BERT scores, indicating that the generated summaries captured the essence and semantic meaning of the source document well. Our lower ROUGE scores indicate fewer direct matches in terms of words or phrases. Overall, a high BERTScore and a low ROUGE score indicated that the generated text was contextually and semantically accurate, while diverging in wording from the source text. This can be explained because we desired to generate summaries for the general public, where the wording should be more simple. Thus, we did not end up replicating exact verbiage from the expert summaries. Our results validated 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 sorted key information into general categories such as Case Information, Parties Involved, and Legal Basis. The second layer refined this categorization using the IRAC (Issue, Rule, Application, Conclusion) framework, a common technique among lawyers. This repeated parsing and combining of information enabled the model to develop a better semantic understanding of the input text while not retaining its original structure.

The CoD prompting technique further structured 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 mitigated 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 organized and verified 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, our framework maintained 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 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 prompting techniques, and exploring solutions in fine-tuning on legal data to enhance the applicability and robustness of our model. Additionally, we will 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

Chalkidis, I.; Garneau, N.; Goanta, C.; Katz, D. M.; and Søgaard, A. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. arXiv:2305.07507.

Felkner, V. K.; Chang, H.-C. H.; Jang, E.; and May, J. 2022. Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models. arXiv:2206.11484.

Niklaus, J.; Matoshi, V.; Stürmer, M.; Chalkidis, I.; and Ho, D. E. 2023. MultiLegalPile: A 689GB Multilingual Legal Corpus. arXiv:2306.02069.

Rickard, E. 2018. Many U.S. Families Faced Civil Legal Issues in 2018. *The Pew Charitable Trusts*.

Shen, Z.; Lo, K.; Yu, L.; Dahlberg, N.; Schlanger, M.; and Downey, D. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. arXiv:2206.10883.

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

Zheng, L.; Guha, N.; Anderson, B. R.; Henderson, P.; and Ho, D. E. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. arXiv:2104.08671.

Östling, A.; Sargeant, H.; Xie, H.; Bull, L.; Terenin, A.; Jonsson, L.; Magnusson, M.; and Steffek, F. 2024. The Cambridge Law Corpus: A Dataset for Legal AI Research. arXiv:2309.12269.