Cross Examine: An Ensemble-based approach to leverage Large Language Models for Legal Text Analytics

Saurav Chowdhury[1], **Suyog Joshi**[2], **Lipika Dey**[2]
[1]Indian Institute of Technology, Jodhpur, India, [2]Ashoka University, India,
Email: chowdhury.4@iitj.ac.in, suyog.joshi_asp25@ashoka.edu.in, lipika.dey@ashoka.edu.in

Abstract

Legal documents are complex in nature, describing a course of argumentative reasoning that is followed to settle a case. Churning through large volumes of legal documents is a daily requirement for a large number of professionals who need access to the information embedded in them. Natural Language Processing(NLP) methods that help in document summarization with key information components, insight extraction and question answering play a crucial role in legal text processing. Most of the existing document analysis systems use supervised machine learning, which require large volumes of annotated training data for every different application and are expensive to build. In this paper we propose a legal text analytics pipeline using Large Language Models (LLMs), which can work with little or no training data. For document summarization, we propose an iterative pipeline using retrieval augmented generation to ensure that the generated text remains contextually relevant. For question answering, we propose a novel ontology-driven ensemble approach similar to cross-examination that exploits questioning and verification principles. A knowledge graph, created with the extracted information, stores the key entities and relationships reflecting the repository content structure. A new dataset is created with Indian court documents related to bail applications for cases filed under POCSO [1] Act. Analysis of insights extracted from the answers reveal patterns of crime and social conditions leading to those crimes, which are important inputs for social scientists as well as legal system.

1 Introduction

Legal language is inherently complex (Marmor, 2014), characterized by formality, precision, and complexity along with use of specialized vocabulary. Legal documents are often lengthy, with intricate reasoning about laws, acts, clauses, and provisions, with redundancy and repetition, as is necessary in the legal domain. Legal professionals, who have to wade through large volumes of legal text daily, therefore look for text processing tools that can help them in searching through the documents, and retrieve relevant information efficiently. Insights extracted from large collections of legal documents benefit different stakeholders like legal practitioners, clients, social scientists as well as law makers. Consequently legal document summarization, sentence / paragraph labeling using classification models and question answering from legal documents have been popular applications of Natural Language Processing(NLP) (Deroy et al., 2021), (Bhattacharya et al., 2021).

Advances in the area of Natural Language Processing(NLP) have inspired a large volume of work in the area of legal analytics across the world. Legal document summarization, both extractive and abstractive, in different languages have been reported all over the world. Legal data analytics is a relatively new area, but gaining rapid popularity. It may be noted that most of the earlier systems were developed using supervised machine learning methods, where the models were trained with large volumes of carefully annotated data, obtaining which is prohibitively expensive. Besides, each system catered to a specific use case, for which it was trained, therefore requiring substantial rework for extension to other legal domains or jurisdictions. With the evolution of Large Language Models (LLMs)(Topsakal and Akinci, 2023), trained on massive volumes of heterogeneous data from a wide variety of sources, the domain of text processing is seeing a paradigm shift. Applying them for legal text analytics is also being explored. However, one of the key challenges of working with these models is the restricted context length on which

[1]Protection of Children from Sexual Offences Act (POCSO), 2012 : Indian law to protect children from sexual offences

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 194–204

they can work, which is far less than a standard legal document. Another challenge stems from the fact that LLMs are known to hallucinate, or generate text that may not be contextually relevant or correct, which is also not quite acceptable for legal text processing tasks. The current work was motivated by these challenges and sought to explore whether these challenges can be overcome or bypassed, thereby easing the tasks of legal document summarization and question answering for analytical insight generation.

This paper proposes an LLM-driven legal text processing pipeline that first generates contextually relevant summaries from long court proceedings, and subsequently uses the summaries to extract legal information to build a knowledge graph from a legal text repository. The knowledge graph is designed following a legal ontology design that stores the core legal concepts and relationships among them. The summarization process uses a "summary of summaries" approach along with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and LLMs (Topsakal and Akinci, 2023). Legal information extraction from the summaries is realized using an ontology-driven LLM-powered question-answering system that employs a novel ensemble-based approach. The ensemble approach was motivated by the method of "cross examination" that is used during legal trials to challenge the credibility and reliability of a witness's testimony, uncover inconsistencies, and present an alternative interpretation of the facts. This process is crucial in legal systems to ensure a thorough and fair examination of the evidence. In the current context, the term "cross-examination" is used metaphorically and the key idea is to follow a multi-pronged information retrieval over two phases. In the first phase, for a given type of information to be extracted from a legal document, a set of paraphrased questions are formulated using ontology definitions of concepts and relations. In the second phase, a set of verification questions are formulated with the retrieved answers. The above design of the ensemble of questions is aimed at establishing the validity of LLM-generated answers by analyzing them from multiple perspectives. Answer to a single question can be compared with its counterparts generated by the paraphrases, and also verified against the original document before it is accepted. Our experiments also show that the quality of answers obtained from the summaries are often better than those obtained from the original document. Based on the fact that LLMs work better on shorter contexts than longer ones, this result is as expected.

Besides the presentation of a novel workflow, contribution of this work also lies in creating a new dataset consisting of Indian court proceedings related to Protection of Children from Sexual Offences (POCSO) Act, 2012. The POCSO Act, 2012 is a gender-neutral law that was passed to protect children from offences including Sexual Assault, harassment, threatening and child pornography. Analyzing cases under the purview of POCSO Act can unravel insights about the nature of the crimes, social and economic circumstances related to the event of crime, existence of bias, if any, in the judicial system, and also uncover the dynamics of how such cases proceed (Damodharan et al., 2021). The knowledge graph contains all details extracted from the court proceedings related to bail applications from the accused. Experiments and evaluations on this collection show promising and interesting results, thereby establishing the feasibility of the proposed methods in setting up a legal text analytics pipeline in a completely unsupervised way. The repository is publicly available.

The rest of the paper is organized as follows. Section 2 presents a review of earlier work in the area of legal text analytics. Section 3 presents the details of the proposed pipeline for summary generation and question answering from the summaries. Section 4 presents details of dataset creation, experimentation and results obtained. Section 5 discusses the experiments and the results. Section 6 concludes with a discussion of some limitations, which are addressed in Section 7, followed by the dataset link in Section 8.

2 Legal Text Analytics - review of earlier work

Automated summarization of legal documents has been an active area of research for quite some time. Two types of summaries are prevalent—extractive and abstractive. Extractive summaries contain a subset of sentences identified as important from the original document. Abstractive summaries may contain new words and sentences, which are strewn together to convey the original content with a reduced size but without losing the original meaning. In Bhattacharya et al. (2021), authors have emphasized the use of extractive summaries for legal documents to ensure that the characteristic of legal language is retained in the summary.

Abstractive summarization stores the essence of a document, but does not preserve exact sentential structures. Masked language models like T5, BART, etc., were found to work well for summarization. In Zmiycharov et al. (2021), a T5-based abstractive summary generation model was proposed for EU legal documents. In Elaraby and Litman (2022), a BART-based model, that could capture the argumentative structure of legal documents by integrating argument role labeling into the summarization process, was proposed. In Feijo and Moreira (2023) proposed an abstractive text summarization approach using an encoder-decoder architecture. Most of the above approaches required a large training corpus to train the models. The models were not transferable and hence not usable in a context other than for which they were designed. To overcome the challenges of high-quality training data, a transfer learning based approach that exploits extractive and abstractive techniques simultaneously, was proposed in Moro et al. (2023). Though Large Language Models (LLMs) can summarize content pretty well and are known to work with little or no supervision, legal document summarizing still pose a challenge since these are very large and often do not fit into acceptable context lengths.

Legal question answering and text analytics beyond summarization is emerging as an important area. Martinez-Gil (2023) presents results of a quantitative and qualitative survey carried out to document the existing challenges in the area, the primary one being the fact that the task is time-consuming and error-prone. Guha et al. (2024) reports a study on the adoption of Large Language Models(LLMs) by the legal community. They present a collaboratively constructed legal reasoning benchmark consisting of 162 tasks covering six different types of legal reasoning called Legal Bench.

Summarizing documents from Indian court proceedings using NLP techniques is a relatively less explored area. Recently, platforms like SCC Online (SCC Online, 2024), Manupatra (Manupatra, 2024), and Indian Kanoon (Kanoon, 2024) have started hosting vast repositories of digitized court proceedings with advanced search capabilities. Bhattacharya et al. (2023) presents transformer-based models for rhetorical role labeling to assign labels such as Fact, Argument, Final Judgement, etc., to sentences of a court case document. Quevedo et al. (2023) presents a detailed study on the readiness of general-purpose LLMs for abstractive summarization of legal documents. They propose a human-in-the-loop approach for obtaining functional summaries with LLMs.

3 Ontology-driven framework for Legal Document Summarization and Analytics using Large Language Models

Though Large Language Models (LLMs), trained over very large repositories are known to be good for general - purpose language generation tasks like summarization or question answering, performance of similar tasks over specialized domains can greatly benefit from the use of ontologies or knowledge graphs (Agrawal et al., 2024). This is true for all specialized domains like health, climate or legal repositories. A legal ontology is well structured framework that defines the relationships between various legal concepts, entities, principles and processes enabling a systematic understanding of law and legal domain. Legal documents are an important part of the ontology (Van Engers et al., 2008). While some legal documents store information about the legal processes, statutes etc. and are more permanent in nature, legal documents arising out of legal proceedings contain details about a specific case, referring to other legal concepts. Each legal proceeding usually gives rise to a number of legal documents of different types. Figure 1 presents a portion of the legal ontology used for our work. This ontology is created from concepts and relations presented in Leone et al. (2020).

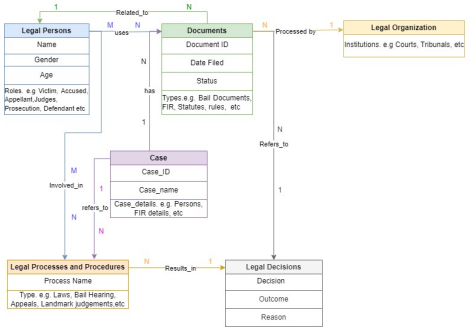


Figure 1: Legal Ontology Diagram

The ontology depicts the interplay between legal professionals, persons, documents, organisations, principles and process. For instance, it shows that accused, victim, lawyer, judge all are "Legal Persons", with various properties like gender, age etc. Legal documents encompass a large category of documents like case files, bail applications, court

proceedings, judgements, contracts etc. Each has its own characteristic (not shown in the figure). The ontology plays a key role in information extraction from document summaries, explained in section 3.2. But before that, in the next subsection we present the mechanism for creating document summaries from legal documents using LLMs. Though the ontology is not used for the task of summarization, it plays a key role in its evaluation, wherein it provides the list of key concepts and relations that should be present in the summary.

3.1 Legal Document Summarization using summary of summaries

Legal document summarization aims to generate shorter versions of long documents retaining crucial legal information components like the judgements, citations, bills under process, acts and laws etc. Long documents are broken into fixed size chunks, say, D1, D2, ..., Dn. Vector embeddings of each chunk are created and passed on to the LLM for generating summaries using the following prompt - "*summarize the provided text. This is just part of a larger document, so do not add any extra information and narration. Provide details about the victim and accused, including gender, age, legal status (minor or adult), relationship between appellant and accused, familiarity between victim and accused, specific charges, repeat offenses, bail approval, final judicial decision, rationale provided by the judge, and relevant legal principles or precedents referenced, do not write anything extra, just reduce the words: {text}*." The chunk summaries are concatenated to create an intermediate document, which is passed on as context to the LLM for a second time with the same prompt to generate the final summary. Figure 2 for an architectural diagram of the approach.

This approach has multiple advantages. In the first pass, it allows the query to focus on each part of the document and include the contextually relevant parts in its summary. In the second pass, it eliminates redundancies. The two passes ensure that if a piece of relevant information is present in multiple chunks but in different contexts representing different perspectives of its use, these are retained in each chunk-summary and also in the final. This holds for citations which are references to past cases, and are often found in both the arguments and judgment sections, but may or may not be used from the same perspective. On the contrary, a piece of information like ones that describe the accused or circumstances of crime, are repeated in different chunks to emphasize on the same truth. These may be retained in the individual chunk summaries in the first pass, but multiple occurrences are eliminated from the final summary.

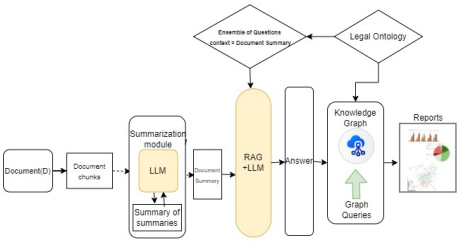


Figure 2: Architecture

3.2 Ontology-driven Information Extraction from Legal documents - an Ensemble-based approach

Retrieval-Augmented Generation (RAG) is a mechanism to improve the performance of LLMs over longer contexts (Fan et al., 2024), by combining the powers of information retrieval and generative models. In this framework, long documents are partitioned into smaller units like sentences or paragraphs, which are converted into text vectors using an embedding language model. The vectors are stored in a knowledge library, from which components relevant to a given query are retrieved and passed on to the generative models for answer generation. RAG based frameworks are gaining popularity as they help generate more accurate, informed and contextually relevant outputs from local repositories.

We now present how the RAG framework is used along with the "cross-examination" motivated ensemble approach to generate answers from legal documents using LLMs. For question answering with LLMs, a key problem that needs to be addressed is to obtain some assurance about the quality of answers, especially since these models are known to generate out-of-context answers, which are sometimes outright wrong. To address this issue we propose the idea of creating an ensemble of legal questions in a controlled manner, using the legal ontology. Lawyers often examine witnesses by paraphrasing an earlier question. Rephrasing a questions helps test the consistency of a witness's testimony. Lawyers also pose clarification questions, wherein questions are formulated with the answers, and the witness has to verify its truth. This is used as a tool to reconfirm or disprove answers

given by the witness. The proposed ensemble design is motivated by this idea of cross-examination.

It is assumed that the information components to be extracted from the legal documents are part of the legal ontology presented earlier. Thus each component not only has a definition, but is known to be constrained by its relationships with other concepts. For a given entity e, a set of questions $Q_p(e)$, which are paraphrases of each other, are generated using its definition from the ontology and an LLM prompt. The questions can also be generated using a mixture of human-paraphrasing and LLM-paraphrasing. Each question in $Q_p(e)$ is passed on to the LLM again for generating answers, this time with a specific legal document summary as the context. Further to paraphrasing, an added layer of confidence in the answers is derived through verification. This is done by creating a second set of questions, denoted by $Q_v(e)$, whose purpose is to verify the answers generated for questions in $Q_p(e)$. Let a_i denote the answer to a question $q_i \in Q_p(e)$. A verification question q_i^v is created for q_i to verify whether the answer a_i is supported by the document. Verification questions are designed as a prompt that will generate either "Yes" or "No" as answer, when the question is passed on to the LLM along with the document as context. A verification question typically looks like *Given the following context, is* a_i *true?* or *"Given the following context, does* a_i *follow* q_i*?"*. For a single question q, thus a multitude of answers is generated with multiple accesses to the LLM.

The above steps are followed for each information component to generate an ensemble of questions for each one of them. The legal document summaries generated in the earlier step, are now passed as context one at a time, along with the questions. Since the questions in $Q_p(q)$ ask the same thing in different ways, it is expected that if all the answers are similar, and each are verified to be correct by the second set of questions, then the answer is right. However, this does not provide an absolute guarantee. Answers to legal questions can be either objective or subjective. Asking about the appellant's name or gender, the police station under which a crime event was registered etc. are examples of objective answers. Subjective answers do not call for a fixed word composition, but need to convey the right sense. The measures defined below are used to determine similarity of two answers:

- ROUGE-L score: This is used to compute the similarity of two texts based on the Longest Common Subsequence (LCS) shared between them. LCS is the longest sequence of words that appear in both the candidate and reference summaries, while keeping the order of the words intact.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) Score: This metric is used for evaluating text generation. It considers synonyms, stemming (matching word roots) and paraphrases while computing similarity. This score reflects semantic similarities that are not captured by ROUGE-L. It calculates precision and recall, giving more weight to recall.
- BERT Cosine-similarity: This measures the cosine similarity of two texts based on their BERT embeddings, capturing the semantic similarity between the two texts.

Based on the definitions, it can be seen that ROUGE-L and METEOR help in verifying whether the expected key concepts are present in a machine-generated summary, while BERT cosine similarity measure captures the semantic similarity of two texts. While question answering is expected to work similarly whether it is done for full documents or document summaries, experiments show that the performance of the process is better for summaries. This could be due to the fact that summaries have almost no repetitions and hence provide easy context to the LLMs. Table 1 shows examples of paraphrased questions and sample answers along with verification questions and answers. The verification questions are automatically picked up from a set of templates curated for the purpose along with the corresponding text from answers returned by the LLM. It can be seen that, while the first and third answers overlap, the second one is not useful. The verification question also proves this.

3.3 Knowledge Graph generation and graph analytics

The extracted information from answers obtained from the questions discussed in Section 3.2 are used to populate a knowledge graph, that follows the legal ontology design presented earlier. The nodes in the knowledge graph represent different kinds of legal entities like case documents as well

Table 1: QA - Sample question and answers with Primary, Paraphrased and Verification questions

Primary Question	Answer	Verification Question	Ans.
Which legal principle, landmark judgment, or precedent was adhered to in the decision?	The Special Court designated under the Protection of Children from Sexual Offences (POCSO) Act has the exclusive power to deal with applications under Section 438 of the Code of Criminal Procedure (Cr.P.C.) related to offenses under the POCSO Act, even before registering an FIR.	Is *answer to question* correct from given context?	Yes
What guiding legal principle, significant case law, or past judicial precedent influenced the document?	Not available	Is it true that the legal principle, significant case law, or past judicial precedent influenced the ruling are *answer to question*?	No
Which principles, landmark judgments, or legal precedents are mentioned as reference points in the document?	Section 438 of Cr.P.C.	Is *answer to question* mentioned as legal principle, significant case law, or past judicial precedent which served as a reference point?	Yes

as laws and sections cited, names of defendants and accused entities along with their properties like age and gender, if retrieved. A set of people-people relations that are not a part of the ontology design and added to the knowledge graph based on the knowledge graph answers. Since LLM generated answers are rather verbose, and not fixed in nature, a named entity extractor is first applied to extract the legal entities from the answers. The extracted entities are then resolved document-wise, using the methods presented in (Kalamkar et al., 2022). For example, a particular statute may be referred to in a legal document multiple times, sometimes with its full name like *Indian Penal Code* and sometimes as *IPC*. A second level of resolution is needed to resolve the entity mentions across the documents, since only one instance of a named entity should be ideally retained in the knowledge graph. For inter-document entity resolutions, we apply a clustering algorithm that uses locality sensitive hashing (LSH) to group similar strings together. Querying a knowledge graph thereafter yields interesting insights about how cases, people or organizations, statutes etc. may be linked to each other.

4 Dataset Creation

We now present the details of the dataset that has been created for this work. This dataset was created keeping in mind an important application of legal text mining, namely analysis of crimes against children. According to National Crime Records Bureau (NCRB), India, 43.44% of POCSO cases end in no convictions due to lack of evidence (Nigudkar et al., 2023). It was also mentioned that only in about 6% of the cases involved an unfamiliar accused and victim pair. In almost 23% of cases, the victim and accused are known to each other, which includes an approximate estimate of 4% of cases where the accused is a family member. According to the NCRB report of 2022, out of 38,444 cases analyzed, 414 or 1% of the cases involved male victims, while the rest involved female victims. The results stated above were manually curated, and have not been updated for last three years. We belive that with proposed mechanisms, one can do these kind of analysis regularly in an automated way. To check the applicability and validity of the proposed framework for insight generation, a repository of 50 POCSO bail applications filed after 2020 has been created. These were collected from two sites eCourts India (2024) and Kanoon (2024). The second site also contains human-generated summaries for these applications, along with human annotations for various sections of the documents, which have been used for evaluation purposes. The questions were designed to extract insights like those mentioned by NCRB. The full list of questions is presented in Appendix 1 (10.1).

5 Experiments and Results

All experiments were done using the Langchain (Tian et al., 2023), (Muludi et al., 2024) platform which facilitates Retrieval-Augmented Generation. Results are provided for LLAMA 2 (Touvron, 2023) and GPT-3.5-Turbo (OpenAI, 2023). The details of evaluation and results obtained are presented in the following subsections. All experiments were run thrice and average results are presented.

5.1 Evaluation of LLM-generated summaries

For evaluating the summaries generated, two experiments were conducted. In the first experiment, GPT-3.5 and LLama 2 were deployed to generate summaries from the whole document. Though GPT 3.5 could generate the summaries, Llama

2 failed to generate summaries from the whole document. In the second experiment, the summaries were obtained using the proposed summary of summaries approach, using document chunks. The summaries were compared with original summaries and notes available in Kanoon (2024) using ROUGE-L, METEOR and BERT similarity scores. Table 2 shows the results obtained for both the experiments. Clearly summaries generated using the proposed summary of summaries approach fared better, when compared to human summaries. We conclude that the restricted context of a chunk helps it to pick up more relevant material for the final summary, than when it works on the entire document at one go. While the higher ROUGE-L scores indicate higher presence of actual legal terms in the second set, the higher BERT similarity scores indicate higher semantic similarity and the METEOR scores indicate lexical matching, semantic meaning, and content coverage. It is also observed that the free Llama2 model performs slightly better than the subscription based model GTP-3.5 Turbo. Compression ratio for a summary is obtained as

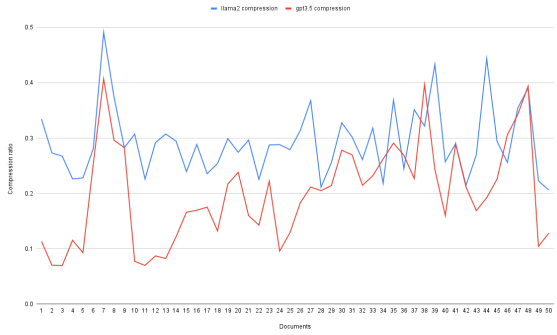


Figure 3: Compression Ratio: Number of tokens in Summary vs Original Document

the ratio of number of tokens in summary against the original document. GPT 3.5 Turbo consistently generates briefer summaries than LLAMA2. Figure 3 presents the compression ratio for the entire set for both the LLMs. The correlation between the results is 0.48, which is quite high.

Manual Assessment of Quality of Summaries: One of the key concerns expressed by legal professionals about automatically generated legal document summaries is the loss of rigour that is a characteristic of legal language. Since the rigor actually stems from the redundancy and repetition, which are dispensed off in a machine-generated, the summaries cannot be used as legal documents themselves, but can help in quick assimilation of content

Table 2: Average Scores for Generated Summaries from 50 Bail Documents for POCSO cases

Context	LLM	Rouge-L	BERT	METEOR
Full Document	GPT-3.5 turbo	0.17	0.80	0.24
Full Document	Llama2	-	-	-
Summary of Summaries	GPT-3.5 turbo	0.21	0.83	0.26
Summary of Summaries	LLAMA2	0.26	0.84	0.36

and answer legal questions posed by lawyers while doing their background research. For that purpose the summaries need to be factually correct in terms of all entities and their roles, cite the correct laws and statutes, be causally correct while reasoning about facts and arguments, and also be readable. Since this is an expert-intensive task, we could obtain expert evaluation for 25 summaries. The ones generated by Llama2 were selected for evaluation. The experts were requested to assign scores between 1 to 5, 1 being the lowest and 5 the highest on the following parameters (i). Correctness of facts (ii). Laws and Statutes (iii). Legal Language (iv). Reasoning correctness. Table 3 shows the average scores obtained.

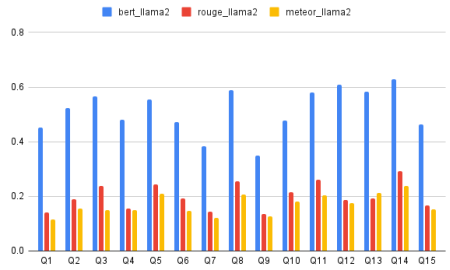
Table 3: Evaluation Parameters and Averages

Parameters	LLAMA 2	GPT-3.5 turbo
Correctness of facts	3.66	3.2
Laws and Statutes	3.66	3.3
Legal Language	2.67	2.9
Reasoning correctness	3.33	3.25

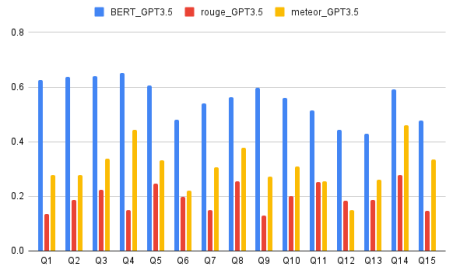
5.2 Evaluation of Question Answering based Information Extraction

We now present evaluation scores for answers to the cross-examination comprising a set of 15 questions, along with their paraphrases and verification questions. For each question a gold-standard human answer was obtained from experts or from Kanoon (2024). The machine-generated answers of the paraphrased questions were compared with the human answers using ROUGE-L, BERT similarity and METEOR (Metric for Evaluation of Translation with Explicit ORdering) scores. Each of these

scores were then multiplied by a factor of 1 or 0, depending on whether the corresponding verification answer was true / false. A weighted average was thereafter computed for each question, for each measure for each document. Figure 4 presents the ROUGE-L, BERT cosine similarity and METEOR scores for this set, averaged for all document summaries. It can be seen that the ROUGE-L scores are fairly consistent across all questions and both the LLMs, and much lower than the corresponding BERT scores. This is expected. However METEOR scores are almost similar in case of answers generated by LLAMA2 but are higher in case of answers generated by GPT 3.5. Figure 4 shows that GPT 3.5 generated answers in general score better than LLAMA2 for most questions, and particularly for those that need inferring, like questions 7 (about relationship between victim and accused), and 9 (whether accused is repeat offender). This also holds for questions 1 to 4, which though appear to be simple, need inferences to be drawn, as these may not be explicitly mentioned in the documents. LLAMA2 does a better job of identifying citations and section numbers etc. It may be surmised that since LLAMA 2 summaries were longer than GPT 3.5 Turbo summaries, they preserved information components better than the later.



A. Scoring LLAMA2-generated answers against human answers



B. Scoring GPT-3.5 Turbo answers against human answers

Figure 4: Comparing LLM-generated Answers with human answers using BERT Similarity, METEOR, and ROUGE-L scores

5.3 Knowledge Graph from Legal Repository: Obtaining Insights from the graph

We have used the Neo4j (Neo4j, Inc., 2023) platform to store and query the knowledge graph generated from the current repository. Figure 5 shows a portion of the knowledge graph with case documents and their references to statutes and laws. For insight generation and analytics we query the graph database using graph query language CYPHER. Besides obtaining the most referred to laws and sections, the most important application of the knowledge graph is to find similar cases, where similar cases are those that might be discussing about similar crimes and hence referring to same or overlapping set of laws. Neo4J identifies similar nodes using a graph based similarity computation, which takes into account structural similarity. Figure 6 shows two such case documents which were inferred as similar. It was found that both these documents refer to similar sets of sections awarded for *gang rape*.

Among other insights found from the answers, we report that 5 out of the 50 cases, i.e. 10% of the cases involved male victims, which is higher than the figures reported earlier, and can be investigated further. Only 24% of the cases led to acquittal of the accused. A clique of 10 cases citing *Scheduled Castes and Scheduled Tribes Act*, suggest subjugation of marginalised section. In more than 50% of the cases, reference to Section 29, indicates that unlike other court proceedings which hinge on the accused's innocence till proved guilty, for POCSO cases, it is presumed that the accused has committed the offence, until contrary is proved. Degree analysis reveals that, most of the cases involved heinous crimes falling under Sections 3 and 5 of POCSO which deals with penetrative assault. Around 10 % cases involved handling of child pornography.

6 Conclusion

In this paper, we have explored how LLMs can be leveraged to perform legal text analytics. We have proposed an efficient mechanism to generate summaries for legal documents using LLMs, with no further training. We have also proposed a mechanism to generate a knowledge graph from a repository of case documents, using across-examination

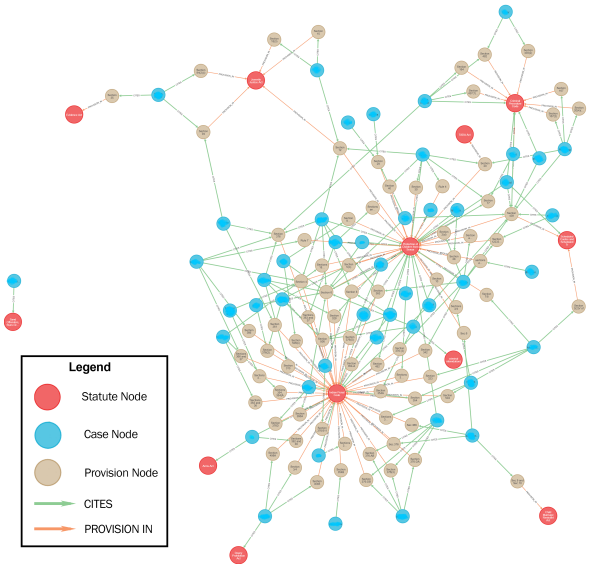


Figure 5: POCSO:citation Knowledge Graph

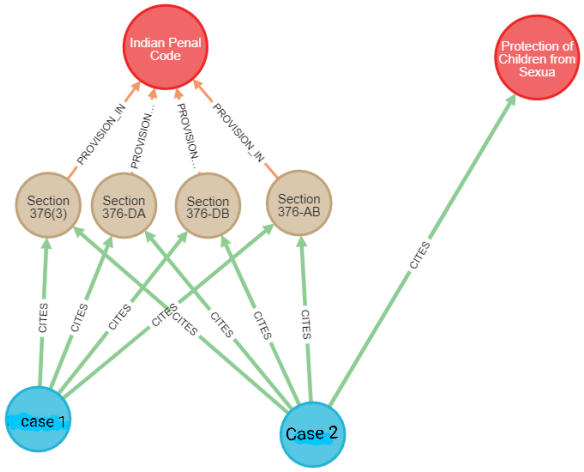


Figure 6: POCSO:citation Knowledge Graph

like technique of posing a set of questions and cross-verifying the answers. Going ahead, we intend to build a completely automated pipeline for legal document analytics and summarization. The dataset has been shared in Section 8. Along with building a large knowledge base, one aspect of research will be focused on automated evaluation of LLM-generated content. Validating the answers through external causal frameworks is also being explored.

7 Limitations

The novel approaches have been tested on a small dataset, so this needs to be thoroughly evaluated on a larger dataset.Going forward we plan to expand the dataset. Further, we plan to implement the pipeline on Large Language Models (LLMs) with larger context window size. Better evaluation scores and methods need to be evolved for legal text analytics.

8 Dataset- Link

The dataset can be found at this GitHub link.

9 Ethics Statement

Our research adheres to the ethical standards, ensuring data privacy by anonymizing all collected data and conducting a thorough bias analysis to mitigate potential harms. All data and dataset used and created are adopted from publicly available resources and adhering to the usage policy. All research paper, journals, websites used in the paper have been duly cited.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. pages 22–31.

Dinakaran Damodharan, Lakshmi Sravanti, R KiragasuruMadegowda, and John Vijay Sagar. 2021. The protection of children from sexual offences (pocso) act, 2012. *Forensic Psychiatry In India*, 66.

Aniket Deroy, Paheli Bhattacharya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. *An Analytical Study of Algorithmic and Expert Summaries of Legal Cases*.

eCourts India. 2024. ecourts - services for indian judiciary. Accessed: October 10, 2024.

Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. *arXiv preprint arXiv:2209.01650*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards

retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Diego de Vargas Feijo and Viviane P Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1):91–113.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, and Diego Zambrano. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

Indian Kanoon. 2024. Indian kanoon - search engine for indian law. Accessed: October 10, 2024.

Valentina Leone, Luigi Di Caro, and Serena Villata. 2020. Taking stock of legal ontologies: a feature-based comparative analysis. *Artificial Intelligence and Law*, 28(2):207–235.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Manupatra. 2024. Manupatra - legal search engine. Accessed: 2024-05-31.

Andrei Marmor. 2014. *The language of law*. OUP Oxford.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. 2023. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, pages 1–29.

Kurnia Muludi, Kaira Milani Fitria, and Joko Triloka. 2024. Retrieval-augmented generation approach: Document question answering using large language model. *International Journal of Advanced Computer Science & Applications*, 15(3).

Neo4j, Inc. 2023. Neo4j. Accessed: 2024-10-14.

Dr. Mohua Nigudkar, Dr. Upneet Lalli, and Soledad Herrero. 2023. Svp national police academy journal june, 2023 vol. lxxii, no. 1. *SVP National Police Academy Journal June, 2023 Vol. LXXII, No. 1*, page 246.

OpenAI. 2023. Gpt-3.5 turbo. `https://openai.com/`. Accessed: 2023-06-16.

Ernesto Quevedo, Tomas Cerny, Alejandro Rodriguez, Pablo Rivas, Jorge Yero, Korn Sooksatra, Alibek Zhakubayev, and Davide Taibi. 2023. Legal natural language processing from 2015-2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access*.

SCC Online. 2024. Scc online - comprehensive legal research. Accessed: 2024-05-31.

Ying Tian, Tianyu Shi, Jerry Gao, and Luheng He. 2023. Langchain: A universal api for integrating language models. Accessed: 2024-06-16.

Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056.

Hugo Touvron. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Tom Van Engers, Alexander Boer, Joost Breuker, André Valente, and Radboud Winkels. 2008. Ontologies in the legal domain. *Digital Government: E-Government Research, Case Studies, and Implementation*, pages 233–261.

Valentin Zmiycharov, Milen Chechev, Gergana Lazarova, Todor Tsonkov, and Ivan Koychev. 2021. A comparative study on abstractive and extractive approaches in summarization of european legislation documents. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1645–1651.

10 Appendix

10.1 Appendix 1: Primary Questions along with two paraphrases of each.

Q1.1 What is the gender of the victim?

Q1.2 Please find the gender of the victim.

Q1.3 What gender does the victim identify as?

Q2.1 Is the victim a minor or not minor?

Q2.2 Please mention if the victim is legally considered a minor or not minor under the age of majority.

Q2.3 Does the victim's age classify them as being under the legal age of adulthood or not adult?

Q3.1 What is the gender of the accused?

Q3.2 Please find the gender of the accused.

Q3.3 What gender does the accused identify as?

Q4.1 Is the accused a minor or not minor?

Q4.2 Please mention if the accused is legally considered a minor or not minor under the age of majority.

Q4.3 Does the accused's age classify them as being under the legal age of adulthood?

Q5.1 Who filed the bail application?

Q5.2 Regarding the bail application, can you mention the name appellant?

Q5.3 Please mention the name who initiated the process of filing the bail application?

Q6.1 How is the appellant related to the accused?

Q6.2 Please mention the nature of the relationship between the appellant and the accused?

Q6.3 Please provide details on the connection between the appellant and the accused?

Q7.1 Was the accused known to the victim?

Q7.2 Did the victim have any prior acquaintance with the accused?

Q7.3 Was there any pre-existing familiarity between the victim and the accused?

Q8.1 Under which sections have the accused been booked?

Q8.2 Under what legal provisions was the accused charged?

Q8.3 What are the specific sections of the law under which the accused was implicated?

Q9.1 Has the accused committed repeat offense?

Q9.2 Has the accused engaged in a repeated offense?

Q9.3 Did the accused commit the same offense again?

Q10.1 Was bail granted to the accused?

Q10.2 Did the accused receive bail approval?

Q10.3 Was bail approval given to the accused?

Q11.1 What was the final decision of the judge for this application?

Q11.2 What was the final verdict of the judge for the case?

Q11.3 Did the judge finally grant bail to the accused for the case?

Q12.1 What were the judge's reasons for the decision?

Q12.2 What rationale did the judge provide for the verdict?

Q12.3 What were the judge's justifications for the ruling?

Q13.1 Which legal principle, landmark judgment, or precedent was adhered to in the decision?

Q13.2 What guiding legal principle, significant case law, or past judicial precedent influenced the document?

Q13.3 Which principles, landmark judgments, or legal precedents are mentioned as reference points in the document?

Q14.1 What jurisdiction does the case fall under?

Q14.2 In which jurisdiction does the case fall?

Q14.3 Under which jurisdiction does the case lie?

Q15.1 In which police station was the case reported?

Q15.2 At which police station was the case reported?

Q15.3 Where was the case reported to the police?