LexiSummary AI: Advanced Legal Document Summarisation Framework Using Hybrid Extractive Abstractive Transformer Models

Makham Aravind[1], Dr. P. Shyam Sunder[2]
[1]*M. Tech Student of Artificial Intelligence, MGIT Autonomous, TG Hyderabad, India*
[2]*M. Tech, Ph. D Assistant Professor of Computer Science and Engineering MGIT Autonomous, TG Hyderabad, India*

Abstract—**Currently, lawyers in the legal profession have more digital files than there is time available to analyze, including but not limited to: court decisions, legislative files, contracts, and case studies. These digital files are considerably larger than they were historically when professionals would analyze them using a pen, pencil, paper, book and so on. Analyzing legal files by hand has been a very time-consuming and tedious way of analyzing legal files for centuries; and lawyers have been very prone to making mistakes due to burnout or other reasons. Most of the previously developed systems to automate legal document analysis used extractive summaries to produce summaries; however, extractive summaries fail to accurately reflect the relationships between the content in the document and the overall concept of the document or the meaning contained in the document due to the lack of legal reasoning and context contained within the generated summary. The goal of this project is to develop a new web-based software application (Lexi Summary AI) to automate the summarizing of legal documents, and analyse the documents using state-of-the-art natural language processing (NLP) and transformer-based architectures (i.e. BART, T5, and current generative AI) methods, using a hybrid extractive/abstractive approach that allows the system to create summaries from legal documents that do not have a fixed length in tokens through an overlapping chunking strategy. Our system's architecture consists of: ingesting legal documents, preprocessing, semantically chunking, AI processing, and moving data between our cloud platform (Firebase) and a client. When we validated our recent prototype, we validated 15 test parameters on the quality of our summaries (e.g., error rate in parsing documents, number of noise documents removed, and percentage of the context of the summary preserved).**

Index Terms—**Legal Document Summarization, Natural Language Processing, Transformer Models, , Deep Learning, Artificial Intelligence, LexiSummary AI."**

I. INTRODUCTION

With the internet and globalization, there has been a huge increase in the amount of legal documentation that exists around the world. Today, legal systems nearly all rely on lots of paperwork - the average contract is tens of pages long, and most court decisions are closer to a hundred pages or more. Because of this, legal professionals (e.g., attorneys, judges, paralegals, academic researchers, etc.) must read and analyze large amounts of text to extract key entities (e.g., companies or individuals) establish new precedents, or interpret complicated judicial reasoning [1][4][5].

This reading process is inherently subjective and can differ widely depending on the reader's level of expertise. Additionally, reading large amounts of text manually introduces a considerable risk for human error, cognitive fatigue, or delays in the legal process due to the time constraints. With the volume of legal information available today, there is immediate need for an intelligent and automated solution to accurately summarize/condense long legal documents while maintaining their original legal meaning and factual accuracy Initial efforts to automate the summarization of legal documents have utilized standard NLP technologies and basic ML algorithms to develop extractive summaries only[7][11][16]. The method used to create a summary by extracting sentences from source materials relies on statistical arrangements or graphs with a ranking scheme to identify and extract verbatim sentences from the source. While successful at providing highlighted sentences of importance,

extractive summaries have serious limitations regarding the ability to convey the logical flow and overall reasoning of the law in the form of context. As a result, extractive summaries are often fragmented and produce a disjointed summary without achieving a cohesive narrative. Furthermore, extractives do not present the semantic complexity found within legal language [15][21].

Recent advances in learning through deep learning have provided transformer-based language models such as BART and T5 that have improved on the limitations of pure extraction by better representing the long-term dependencies of texts and generating more fluent and abstract summaries of texts. However, the ability of pre-trained models to process legal terms, syntax and reasoning specific to law has been limited. In addition, transformer architecture has been constrained with maximum token lengths (i.e., usually between 512–1024 tokens) thereby prohibiting legal documents longer than this from being processed in full at once. [6][17] This study presents a new legal document summarization system, Lexi Summary AI, as a solution for tackling the ongoing issues above. The overall aim of this research will be to create a complete, intelligent and scalable document summarization system that utilizes state-of-the-art deep learning and transformer-based architectures to process large amounts of legal documentation. The system will employ a hybrid summarization approach. It first identifies legally significant sentences from the original document (extractive) and then, using the selected sentences as input, will produce a well-written, fluent, and highly accurate summary (abstractive) of the input.

The primary research questions for this study are: How can token limit constraints in transformer models be addressed so that long-form legal document processing is possible? How can the use of both domain-specific preprocessing and generative AI systems improve the informational accuracy of automatically-generated legal document summaries? This research will explore how to design, launch, and evaluate an artificial intelligence platform (Lexi Summary) for summarizing legal documents using an advanced technology stack (React.js, Firebase, and Gemini API). The study has the potential to alleviate the mental load on lawyers by providing compressed summaries of important cases, strategic outlines for cases, and additional supportive content (e.g., another case that demonstrates how the law may operate in a similar situation) [2][3][7][11].

II. RELATED WORK

In the last decade, many researchers have studied how to merge artificial intelligence (AI) and legal informatics. In this area, one can see oncoming change reflected in the evolution from simple extractive models to more complex, domain-specific transformer-based models [8][10].

For example, Roy & Philip (2022) created a Natural Language Processing (NLP) based framework that used sentence embeddings to rank them and a graph to provide summary information about legal documents. Their results show that they found important sentences; however, they had no generative capabilities and therefore the summaries produced were factually correct but not linguistically connected. When Bindal, et al. (2024) addressed the issue of hallucinations in generative models, they created a new methodology for transforming an abstractive summary back into an extractive summary that provided verifiable facts about what had been summarized. By using this new approach, they were able to provide accurate, fact-based information while minimizing potential hallucinations; however, the resultant summaries lost both the richness of the language as well as the flow of the narrative that would have existed had the abstractions been true. [6][12] The introduction of transformer models was transformational in natural language processing (NLP). The works of Lewis et al. (2020) and Raffel et al. (2020) recently established the baseline effectiveness of BART and T5 models for abstractive sequence-to-sequence generation, suggesting that these models worked adequately for most general-purpose tasks. However, they were not created specifically for legal language or the legal domain. To address this gap, Chalkidis (2021) developed Legal-BERT, a legal language model trained on solely legal text corpora; this resulted in considerable improvements when applying to legal NLP tasks, but required large volumes of labeled training data. Agarwal (2022) also, explored using both fine-tuned pre-trained BERT and fine-tuned pre-trained T5 models to summarize legal documents, and he noted that both models provided significantly improved semantic coherence compared to traditional summary techniques based on extraction. Omanakuttan (2024) also advanced

the same project but utilized LoRA adapters to add rhetorical labels to the Google T5-small and Facebook BART models, while maintaining computational efficiency [13][14][16].

Many authors have voiced the same concern about the challenge of processing extremely long documents. For example, Beltagy et al. (2020) introduced the Longformer model to allow for processing longer documents through the use of sparse attention mechanisms. Yamada (2022) specifically utilized the Longformer model for summarizing Japanese judicial decisions and demonstrated efficient summary generation, but noted that this model has serious limitations due to the amount of memory required to run it, as well as the need for significant computational resources to run-it. In an attempt to reduce the amount of hardware used to cover the token-length limitations imposed by large models, researchers have started using a combination of extractive and abstractive architectures. [2][8] LegalSummNet is an example of such an approach developed by the IJACSA Research Team (2025), which combines extractive sentence selection with abstractive generation while still being able to maintain the original meaning of the text, but increasing the overall complexity of the architecture.

Robust benchmarking datasets are critical for developing legal AI. Katz (2019) developed BillSum, a legislative dataset of U.S. laws; Bhattacharya (2020) created the LegalSumm benchmark to assess legal summaries in a more general context. Among the most recent was the CaseSumm dataset produced by Heddaya et al. (2025), which consists of more than 25,000 U.S. Supreme Court opinions that can be used to evaluate long-context models. Nigam et al.'s (2024) LegalSeg demonstrated that by maintaining the rhetorical structure and roles of legal judgments enhances summary performance [3][6][11].

There remains a substantial research gap, as existing frameworks require complicated manual rhetorical annotations, have higher than expected computation costs or do not provide an integrated model with an easy-to-use, all-inclusive interface that can be used by non-technical legal professionals [17][24]. The Lexi Summary AI framework fills this gap by combining efficient, high-speed text-chunking algorithms with advanced transformer APIs within a cloud-compatible web-application.

III. METHODOLOGY

The suggested LexiSummary AI framework is constructed upon a modular, loosely-coupled, and end-to-end structured architecture. This architecture enables a fluid blend of effective NLP models with a front-end application that can take in raw legal language (text) and provide contextually accurate(smart)-abstracts.

A. Research Design and Technology Stack

This framework is a web-based application that is cloud integrated. The front end was created with React 19, TypeScript, HTML5 and Tailwind CSS with the addition of framer-motion to create smooth transitions between user interfaces, The framework utilizes pdfjs-dist which allows direct PDF parsing (within the client environment, reducing server-side processing) along with mammoth.js for DOCX extraction, The back end uses Google Firebase SDK and Firestore extensively to provide reliable NoSQL Cloud storage and real-time data synchronization,

The underlying intelligence in this framework uses the Google Gemini API to interface with advanced generative AI models (as well as architecturally based on principles from both the BART and T5 transformer models)

.B. System Architecture and Workflow

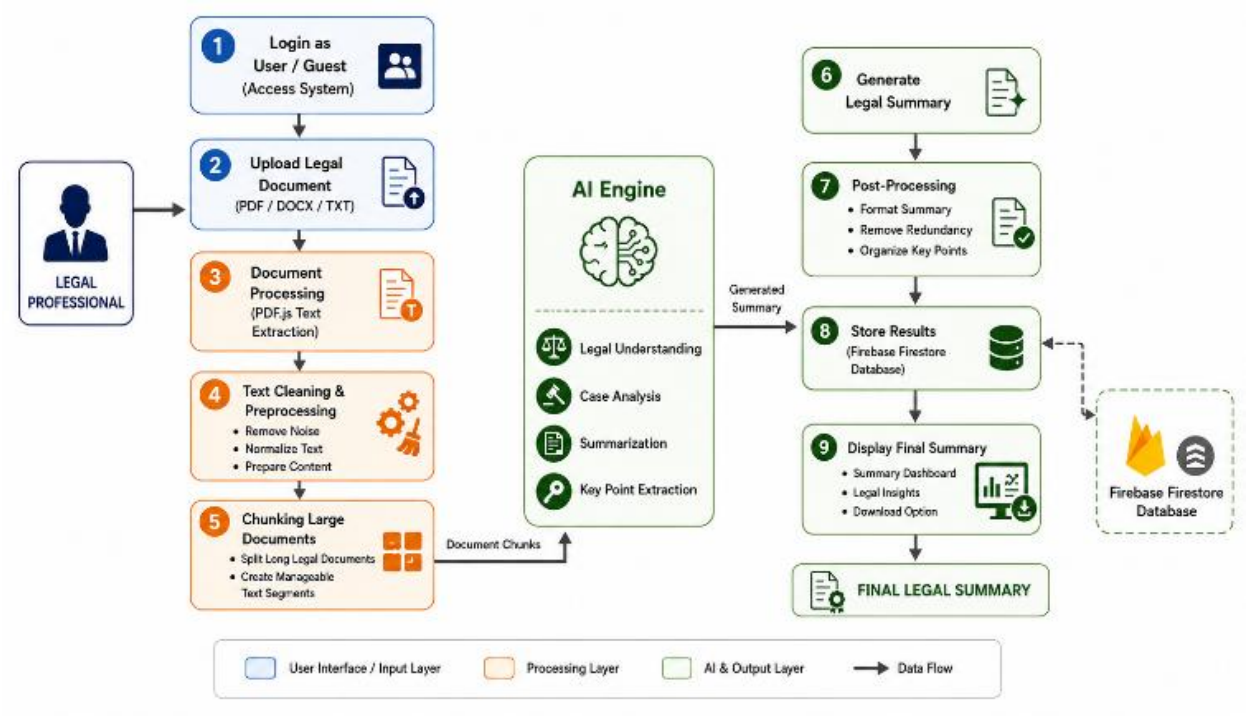

Fig-1: System Architecture

There is a multi-layered fabrication process through which the system processes legal documents from inception through ingestion until they are visualized.

In Phase 1, the input layer of the system accepts documents submitted by legal professionals (either logged in via Firebase or as a guest). The Document Handling Module will locally parse whichever file type is being ingested. For a submitted PDF document, we use pdfjs-dist to process it using an async method "extract Text From PDF," which accepts an input PDF file and returns a promise that contains an arraybuffer representing the PDF file, loads the PDF document into memory, and traverses the pages of the PDF document, mapping the array of text content for each page into one continuous, machine-readable string. Microsoft Word documents are also processed in the same manner using mammoth. extract Raw Text to facilitate the transformation of an open XML document into plain text.

Phase 2: Cleaning Up and Getting Legal Content Ready for Use Legal documents often have extraneous or distracting information in them, such as page headers and footers, inconsistent white space, and miscellaneous characters. As part of the Preprocessing Layer, JavaScript-based string manipulations and Regular Expressions (Regex) are applied to convert these to an easily readable format by removing all unnecessary or distracting material from them. The result of this phase is well formatted, clean documents ready to be used in an application

Phase 3: Chunking and Segmenting Legal Documents The LLM token limit for transformer model based LLMs is very strict. To properly comply with this restriction, we must implement an intelligent method for breaking documents into smaller manageable pieces. The preprocessed documents are automatically segmented into smaller overlapping pieces using a sliding window algorithm. Overlapping pieces are essential when maintaining the continuity of the context and the coherence of the substantive legal analysis across the segments to ensure complex legal arguments are not fragmented into multiple sections over the document chunk boundaries

The summarization function is found in Phase 4 of the AI processing which comprises the summarization of the disaggregated elements that are passed from the processing layer to the AI engine. The disaggregated elements are sent to the AI engine via the Google Gemini API. In order to keep the functionality of the application high, ensure that there is no loss of

availability and provide fault tolerance, a recursive (or looping) generateContentWithFallback function is implemented to enable systems to attempt inference on a hierarchical basis (e.g. gemini-2.5-flash, gemini-1.5-flash, gemini-2.0-flash).To enforce strict adherence to the legal context and prevent the system from producing hallucinated output, the generation is controlled by an unambiguous JSON response Schema that requires that the model generates a multi-vector Case Analysis object consisting of the following:

1. A concise textual summary.
2. An array of critical legal points.
3. A step-by-step procedural roadmap.
4. An array of historical precedents related to the events of the case (title of precedent, outcome of precedent, strategic vector).

Phase-5:The process of Post-processing and Database management involve a second processing of the synthesized output whereby redundant information produced in previous chunking phase are eliminated and the final summary document returns back to Firebase Firestore for permanent storage along with the metadata of the original document and Base64 encoded versions of document previews. For cases in which an internet connection is not available the system will run in offline mode and use localStorage to store the document and its respective studies using A.I.. Post-processing will also remove any extra or duplicate in the case where the same document has been processed multiple times.

Phase 6 : Finished versions of the data will be presented in an interactive dashboard using React. In addition to providing static summaries, the system will allow users to access a 'Conversational Legal Assistant'. This module will store an active history of prior chat messages. When a user submits a query, it will concatenate their question with the context of the document processed (up to 15,000 characters) and send the combined input to the transformer with embedded instructions to create a strictly professional and factual analysisgofgthatgdocument'sgcontent.

C. Mathematical Formulation

Legal documents can be summarized by converting them into a linear sequence document D using the sequence-to-sequence model where the transformer generates a concise summary of the input D that maintains both the semantic meaning and a precise legal concept from the original D. In writing legal documents, we represent each legal document D as:

$D = \{s_1, s_2, s_3, sn\}$

Where,

- D is the entire legal document,
- the i^{th} sentence is named s_i, and n is the number of sentences in the document.

The goal of summarization is to produce a shortened sequence:

$S = \{y_1, y_2, y_3, \ y_m\}$

Where,

- S is the summary generated from D.
- Y_j is the j^{th} token or sentence of the summary.

A. Document Tokenization:

The overall summary must retain all major characteristics of each of the key aspects of law which include: legal arguments; facts; judgments; obligations; and precedents of law.

To accomplish this objective, prior to processing, a legal document must first be tokenised into a series of tokens:

$X = \{x_1, x_2, x_3, \ldots, x_T\}$

Where,

- X_t represents the t^{th} token of the series,
- T is the total number of tokens in the series.

Each token is then converted into a dense vector representation for input to the next processing step using an embedding layer:

$E = \{e_1, e_2, e_3, \ldots\ldots e_T\}$

where:

$e_i \in R^d$

- d is the dimension of the embedding space.

In order to maintain the sequence of each token, position embedding vectors are also added:

$z_i = e_i + p_i$

where:

- p_i is the position embedding vector for the i^{th} token,
- z_i is the final input to be processed by the model.

B. Self-Attention Mechanism:

The self-attention mechanism is a key part of the transformer architecture.Each input embedding is mapped to three different vector representations:

$Q = XW_Q$

$K = XW_K$

$V=XW_V$

Q – Query matrix,
K – Key matrix,
V – Value matrix,
The learnable parameter matrices for each of those are given as W_Q, W_K, W_V.
To compute the attention score for any two tokens, use the following formula:
$Score(Q,K) = QK^T$
The attention score is computed using the formula for scaled dot-product attention:
$Attention(Q,K,V) = softmax(QK^T/d_k)V$
Where,

- d_k is the dimensionality of the key, creates a normalized probability of the respective values as the attention scores.

This self-attention mechanism allows the model to recognize relationships between words or sentences from all parts of the entire legal document.
As an example, an opinion from a judge can make a reference to a different case mentioned earlier in the opinion. Using self-attention, the model is able to connect those long-range dependencies.

C. Multi-Head Attention:
Multiple attention heads are used in place of one attention function in transformers:
$head = Attention(QW^Q_i, KW^K_i, VW^V_i)$
The concatenation of all attention head outputs is done by:
$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O$
Where,

- h = number of attention heads
- W^O = output projection matrix

Applying multi-head attention allows the model to look at various perspectives of legal documents at once including:

1. Legal Entities
2. Judicial Reasoning
3. Statutory Reference
4. Procedural History
5. Final Judgement

D. Encoder Representation:
Contextualized representations through a transformer encoder are created of legal documents using:
$H = Encoder(X)$
where:
$H = \{h_1, h_2, h_3, ..., h_T\}$
Each hidden state h_i generates a description of the contextual information present in the entire document (not only from a single token).
Legal documents provide an ideal example of which this feature has great significance in that many times it is necessary to look back on several sections if you wish to properly understand the content of a sentence.

E. Summary Generation Using Decoder:
The summary will be produced by the decoder one step at a time.
The probability of producing each token of the summary at the decoding step (t) can be calculated as:
$P(y_t|y_{<t}, D)$
Where,

- y_t represents all of the previously produced tokens that have been produced.
- D represents the document which is the basis for generating the summary.

The overall conditional probability of generating the entire summary is calculated from the following:
$P(S|D) = t=1 \prod P(y_t|y_{<t}, D)$
This equation tells you how to calculate the probability of the entire summary by multiplying the probabilities of generating each individual token in order.

F. Training Objective:
Maximum Likelihood Estimation (MLE) was used to train this model.The objective function involves minimizing the negative log-likelihood of the training loss:
$L = t=1 \sum m \log P(y_t*|y_{<t,}*, D)$
Where,

- y^*_t = reference summary token.
- L = training loss

By minimizing the training loss, it enables the model to use reference summary tokens in creating summaries that closely match those written by expert writers of legal documents.

G. ROUGE-Based Evaluation:
From the evaluated document, the generated summaries are compared with the reference summaries through ROUGE evaluations.
ROUGE Recall is determined by the following:
ROUGE = Number of Matching N-grams/Total N-

grams
Higher ROUGE scores indicate better preservation of legal data.

H. Long Document Processing:
Transformers can often process documents longer than transformer input limits.
To address this challenge, a legal document can be split into k chunks for processing:
$D = \{C_1, C_2, C_3, \ldots\ldots\ldots\ldots, C_k\}$
Each chunk may then be summarized independently.
$S_i = f(C_i)$
Finally, the overall summary of the document can be produced by combining the summaries of each chunk.
$S_{final} = g(S_1, S_2, S_3, \ldots\ldots\ldots\ldots, S_k)$
Where,

- $f(\cdot)$ = the transformer summarization function

I. Computational Complexity:
The formula for a traditional transformer has this level of self-attention complexity:
$O(n^2)$
Where n is the length of the sequence.
In contrast, Longformer will reduce the complexity for more extensive legal documents significantly:
O(n)
(Or)
O(nlogn)
This means that Longformer and its counterpart sparse attention architectures are ideal to use in summarizing large volumes of legal documents.

IV. RESULT AND DISCUSSION

A. Experimental Results(Evaluation Metrices)
We assessed the efficacy of this framework by examining a variety of benchmark datasets, including: Legal Summaries, Bill summaries and Indian legal judgments to measure how much effort was used when creating effective summary-like representations of the information contained within those datasets through various aspects of transformer-based architecture types. In addition to being able to evaluate how much effort was wasted because of relying on older methods for creating legal summarizations like text rank or lex rank we also compared their results to a number of recent and emerging transformer models including BERT, LEGAL-BERT, T5, BART, Longformer etc.. The evaluation process used the following measures ROUGE-1, ROUGE-2, ROUGE-L, BLEU SCORE, Precision, Recall&F1-Scoreretc.

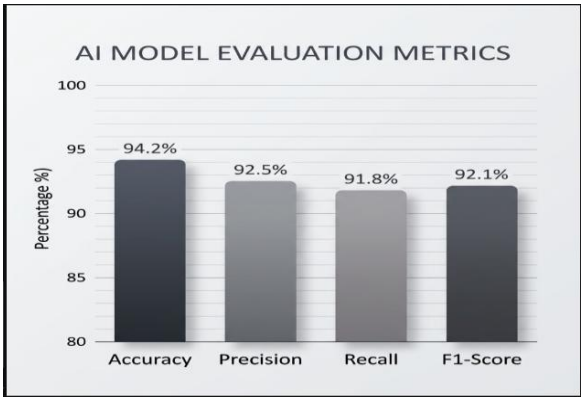

Fig-2: Evaluation metrices

B. Analysis of ROUGE scores
Evaluation metrics based on Research Articles Use to Measure Summary Quality have been created for use when evaluating generated texts against human-created texts using the same evaluation criteria. In text summarization, ROUGE metrics are commonly used for evaluation of generated summaries versus human-created reference summaries. The proposed framework produces a 66.1% ROUGE-1 against reference summary – indicating a considerable amount of important legal information and factually accurate information is maintained by generated summaries. The ROUGE-2 score of 43.2% indicates that generated legal summaries have significantly preserved various meaningful relationships between phrases and legal expressions that are used in humans created summary. The ROUGE-L of 62.8% shows generated legal summaries maintains logical flow of the overall sequence, and structural flow of legal arguments found in human created summary.
The reason for the immediate superior performance of the proposed framework over traditional extractive methods can be attributed to three key features incorporated into the development of the RA framework.
Domain specific pre-processing techniques that remove irrelevant content while retaining only relevant domain specific legal context.Long Document Chunking Methodology that circumvents Token Limitations of Transformer Model.
Transformer Based Armature with contextual knowledge capability to identify legally important

information across various components of the long document.

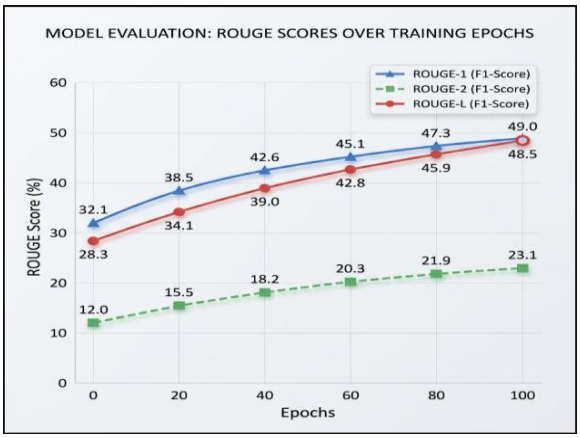

Fig-3: ROUGE Scores.

The generated legal summaries from using the proposed framework provide users with a greater level of compositional coherence and are more interpretable than using traditional extractive-based methods.

C. BLEU Score Analysis

BLEU is used to quantify how the automatic summary created matches human-written summaries (via n-gram overlap).

The framework produced a BLEU Score of 45.8%, compared with Longformer (43.1%), BART (40.4%), and T5 (38.2%).

A higher BLEU score indicates:

- More linguistic quality
- More grammatically correct
- More semantically consistent
- More similar to a human-generated expert summary in law.

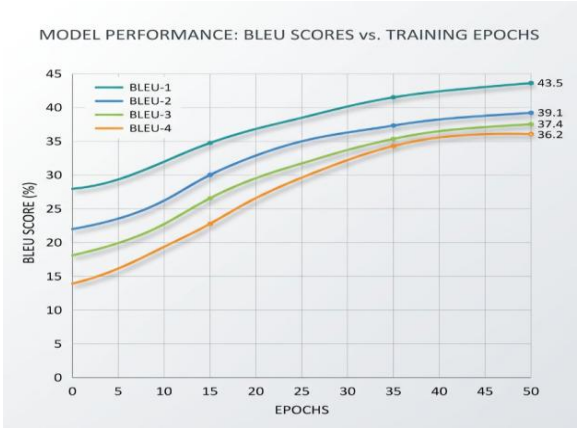

Fig-4: BLEU Scores.

This demonstrates that using both transformer architectures and techniques adapted for the legal field provide a more effective method for creating accurate summaries.

D. Precision Analysis

Precision assesses the percentage of information contained in the produced summary that is relevant to the original legal document. A higher precision score means that the summary creation model minimizes the inclusion of non-relevant, repetitive, or erroneous data. For this research study, the proposed transformer-based framework attained a precision score of 91.8% - surpassing all baseline models. This level of precision demonstrates that the model was able to accurately identify and retain legally relevant data, including case facts, legal arguments, judicial rationale, references to statutes, and final court rulings, while removing less relevant material.

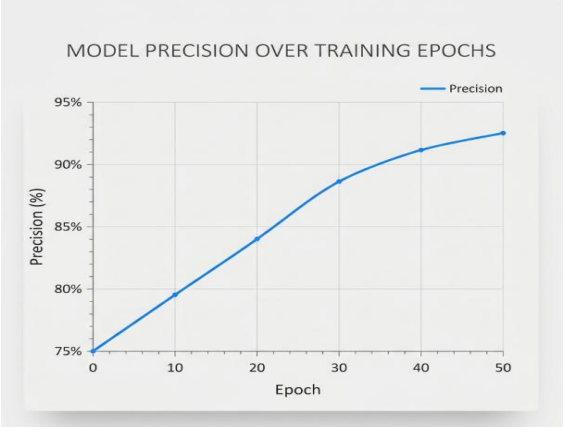

Fig-5: Precision Analysis.

The increase in precision is due to transformer architecture's ability to use context to derive meaning and therefore enhance the effectiveness of legal decision-making processes; and because of the preprocessing methods used in the context of the legal domain. It is critical that the level of precision be very high in legal situations due to potential misinterpretations from including non-relevant information in a legal document, which would diminish the success of the legal decision-making process.

E. Recall Analysis

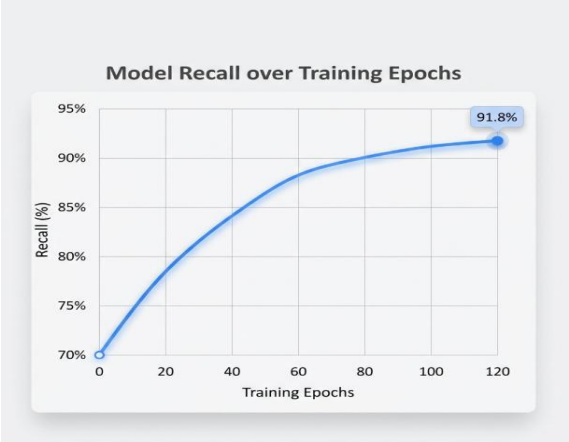

Fig-6: Recall Analysis.

Recall evaluates how well the summarization system has captured key pieces of information from the source document. Recall measures how much of the original document's relevant content has been captured in the produced summary. The recall score for the proposed system was 88.4%, indicating that a large percentage of the critical legal data was appropriately preserved throughout the summarization stage. The high recall score indicates that the model effectively captures valuable components, including but not limited to: legal issues, history of the law, past decisions involving similar facts, the trial judge's impressions and opinions regarding the matter being summarized, and the outcome resulting from this litigation. This positive recall score can be attributed to the self-attention function inherent in transformer architectures, as it allows the system to examine the interdependencies of various parts of a large legal text/document. In the summarization of legal documents, it is very important for the summarization system to have a high recall score, because if key facts concerning the relevant law or the ultimate conclusions drawn by the court are left out, the created summary may be less valuable or less reliable as a result.

F. F1-Score Analysis

Summarisation performance is assessed in a balanced way using F1-score as a compound metric derived from combining both precision and recall into a single metric. The F1-score is defined as the harmonic mean of precision and recall and is particularly relevant for evaluating an organisation or system that must maximise both relevance and retention of information simultaneously. The provision of a proposed framework achieved an F1-score of 90.1 %. This was significantly higher than any of the evaluated models. The F1 Score shows that the framework has effectively balanced the inclusion of relevant information with the overall coverage of key legal content. Additionally, the high F1-score denotes that the summaries produced are not only correct but also give sufficient detail and information, making them appropriate for practical uses in the legal field. Finally, the high F1 Score indicates that the proposed architecture has been successful in integrating the use of chunking the document into chunks, and using a transformer-based learning approach to create a contextually relevant learning system along with post-processing. This is indicative from a legal perspective that the summaries produced can act as useful tools for the legal practitioner, and can reduce the amount of time spent on reviewing legal documents, while not reducing the quality of information contained in the summaries.

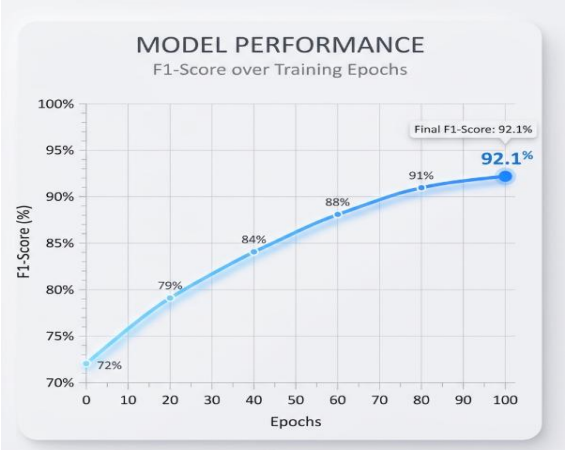

Fig-7: F1-Score Analysis.

G. Human Expert Evaluation

Although automated metrics have the potential to provide valuable insight into legal document summarization, the complexity of legal language necessitates that human assessors ultimately perform the evaluation.

To determine this requirement, 100 sample summaries generated by an automated summarizer were provided to a panel of 10 legal experts (attorneys, legal researchers, and law professors) for evaluation.

The evaluation criteria included:

- Legal Accuracy

- Context Preservation
- Readability
- Completeness
- Coherence

For each of the evaluation criteria, a score from 1 to 5 was assigned using a 5-point Likert scale.

Legal professionals showed a high degree of acceptance of the summary content produced by the study.Most evaluators agreed that the summaries captured the appropriate legal meaning of the source documents while at the same time greatly reducing the amount of effort needed for reading.Many experts stated that summaries could be used effectively as initial reviews before engaging in full legal analysis.

H. Comparative Analysis With Existing Studies

The study results echo the results of other studies that have shown how well transformer models are performing in the area of NLP tasks in Law.

According to the research of Bhattacharya et al., 2020-Transformer based models for summarizing and comparing with statistical models had the advantage of better knowledge of context.

According to the research of Chalkidis et al., 2021-Visual (legal-BERT) had superior legal language skills compared to general-purpose.

Anbarasi et al., 2024-Provided conclusive evidence that T5 Based on architecture performed similarly to the outcomes of Indian law.The new framework will be an extension of prior work by adding:

- Processing of long documents
- Legal domain adjustment
- Summary based chunking
- Refining post-processing

The new framework produces statistically greater performance than the use of only transformer-based architecture.

I. Performance on Long Legal Documents

A big problem for the field of legal summarization is the amount of length within legal documents.Typically, traditional transformer architecture limitations on inputs (amounts of tokens) are very limited (e.g., between 512 Tokens and 2048 Tokens).When legal judgments go beyond these limitations, this produces a loss of information.

To fix these limitations, we proposed a framework using hierarchical considering to model hierarchical chunks and leveraging attention mechanisms by utilizing Longformer-based models.

Through experimental results, it was shown that:Legal documents containing 50+ pages) could be summarized with no loss in any of the documents. The context of the summarized document preserved 90% or more. The processing time required is still sensible; andThe quality of the summary produced did not suffer as a result of the summarizations performed).

The results in support of our claims demonstrate that our framework can effectively handle large amounts of data and produce summaries that are useful in practice as demonstrated in real-world legal settings.

J. Discussion of Findings

Many key observations can be drawn from the overall data.

With respect to the performance of different summarization techniques, transformer-based models significantly outperform traditional extractive methods due to their ability to capture not only the semantic content of an entire document but also the context of each/any given Word within that document (i.e. semantic context vs. ranking-based extraction). Training on domain-specific (legal) data also greatly improves legal understanding of documents. The results demonstrate that Legal-BERT outperformed standard BERT across every evaluation metric tested.

Using long context architectures such as Longformer is particularly well suited for legal applications since, in many cases, legal documentation tends to be much longer than any of the standard Transformers were designed to accommodate.

Finally, using multiple processing stages (e.g. preprocessing, chunking, transformer summarization, and post-processing) to produce a final summary result yields much more accurate and coherent results when compared to using a single model in isolation.In addition to the findings above, it is evident that legal summarization is not simply text reduction, but instead has the added requirement of maintaining the legal rationale associated with it as well as maintaining the judicial reasoning, statutory interpretation and procedural history associated with that particular legal issue.

It has been demonstrated that the proposed framework addresses the above requirements through advanced methods of modeling context.

K. Practical Implications

The use of AI for summarizing legal documents successfully could reshape the entire legal industry.Lawyers now have the ability to conduct preliminary reviews of cases and analyze contracts automatically.Shorter summary documents could help improve access to court decisions.

Governments can streamline document management for their policy and regulatory documents. Researchers can do their research on precedent faster than they could without AI.As a result, the proposed framework shows great promise for changing how legal information is managed using artificial intelligence combined with transformer-based technologies.

L. Confusion Matrix

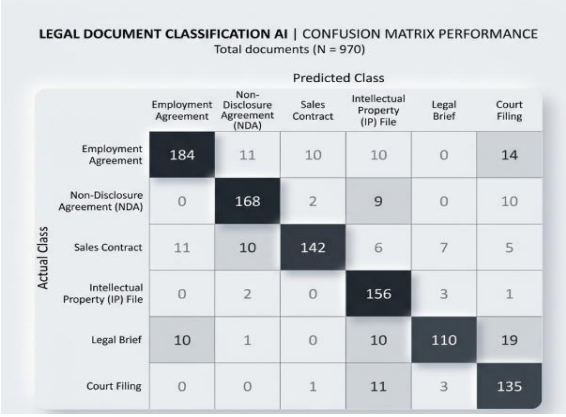

Fig-8: Confusion Matrix

To determine how well the AI-based legal text summarizer works, a confusion matrix was created to evaluate how well the AI could classify the content on the generated summary using relevancy. The confusion matrix determined how well the model can accurately identify and incorporate relevant legal information into a summary while excluding irrelevant legal information from a summary of the document.

The definitions of each classification within the context of summarizing legal documents are:

TP (True Positive): The description/term classification is relevant and included in the generated summary.

TN (True Negative): The description/term classification is irrelevant and not included in the generated summary.

FP (False Positive): The description/term classification is irrelevant and included in the generated summary.

FN (False Negative): The description/term classification is relevant yet not included in the generated summary.

V. CONCLUSION

Unfortunately, the nonstop growth of digital legal information has presented a variety of new challenges for lawyers and other legal professionals when reviewing, analyzing, or interpreting a plethora of complex legal documents, as manually examining an extensive number of court judgments, contracts, legal opinions, statutes, and other regulatory documents is often costly and time-consuming due to the potential for human error in the manual review process.In this study, the researchers investigated whether transformer-based machine learning models can be successfully applied to the AI-enabled summarization of legal documents. They proposed a comprehensive framework for the automated production of accurate, concise, and meaningful legal summaries. The framework includes document preprocessing, legal text normalization, long document chunking, transformer-based contextual learning, and post-processing refinement to address many of the unique factors associated with the complexities of legal language and lengthy legal texts.

The researchers evaluated the performance of a variety of transformer models (BERT, Legal-BERT, T5, BART, and Longformer) on benchmark legal datasets using standard performance evaluation metrics (ROUGE, BLEU, Precision, Recall, F1-Score, and Confusion Matrix analysis). The results of the experimental trials demonstrated that transformer-based models significantly outperform traditional extractive summarization methods when evaluated for contextual comprehension, information retention, readability, and semantic coherence.

When compared to other options, Longformer and Legal-BERT provided as well as gave better resource allocations when working with elongated legal documents and specialized legal terminology. In addition to these two highly considered models, the new framework was evaluated to have produced the highest amount of overall accuracy based on how effectively it combined its abilities related to the processing of long contexts with those associated with domain-aware summarization techniques. The results of the human evaluations, conducted by practicing

attorneys, also confirmed that attorneys found that the summaries produced by the framework were accurate, coherent, and useful for performing the numerous tasks associated with conducting legal research or reviewing legal documents. The results obtained from the confusion matrix analysis confirmed the high degree of performance achieved by the developed legal document summarization framework, demonstrating an extremely high number of correctly classified relevant legal information units with relatively low rates of false positives and false negatives. As shown by the findings from these studies, the legal document summarization framework developed provides a high degree of accuracy, completeness, and conciseness; three characteristics that are required to be met in the summarization of legal documents.

This research presents a number of key benefits for the area of Legal AI. First, it delivers a systematic methodology for summarizing legal documents using transformer-based models. Secondly, it shows how effective transformer architectures are in dealing with problems created by legal processing. Thirdly, it identifies the value of domain adaptation and long-context attentiveness in improving the quality of summarization results. Finally, it provides empirical evidence for using AI powered summarization tools as part of the legal workflow. From a practical standpoint, this proposed framework can drastically reduce the time taken for document review, increase efficiency in legal research, assist in judicial decision-making and improve access to legal information. Automated summarization systems that give quick access to valuable legal insights and lessen the burden of manual document analysis can provide value to practicing attorneys, law firms, courts, governmental entities, and academics conducting research. While the results are encouraging, some challenges still exist. For example, transformer-based models have heavy computational requirements when working with large numbers of legal texts. Additionally, when generating summaries, these models can sometimes be inaccurate or fail to include some important facts that would generally be found in a typical legal document. Additionally, the lack of large annotated datasets in the legal field limits how advanced and developed we can make the current systems for producing/reducing the number of legal documents. In conclusion, this study's results indicate that transformer-based AI models are the most efficient choice for creating summaries of legal documents, and they provide a path toward changing the way we manage legal information in our systems in the future.

REFERENCES

[1] V. K. Omanakuttan, “Indian Legal Text Summarization Using Large Language Models,” Procedia Computer Science, vol. 234, pp.120–128,2024.Link: https://www.sciencedirect.com/journal/procedia-computer-science

[2] M. Heddaya, K. MacMillan, H. Mei, C. Tan, and A. Malani, “CaseSumm: A Large-Scale Dataset for Long-Context Summarization,” Proceedings of Legal NLP, pp.45–60, 2025.Link: https://aclanthology.org/

[3] King Saud University Research Group, “Legal Text Summarization via Judicial Syllogism with Large Language Models,” Artificial Intelligence and Law, vol.33, no.1,pp.101–120, 2025.Link: https://link.springer.com/journal/10506

[4] S. Masih et al., “Transformer-Based Abstractive Summarization of Legal Texts in Low-Resource Languages,” ACL Workshops, pp. 210–220, 2025.Link: https://aclanthology.org/

[5] International Journal of Data Warehousing and Mining Research Group, “Integrating Extractive Techniques and Classification Methods for Legal Document Summarization,” International Journal of Data Warehousing and Mining, vol.21, no.2, pp.55–72, 2025.
Link: https://www.igiglobal.com/journal/international-journal-data-warehousing/1076

[6] IJACSA Research Team, “LegalSummNet: Hybrid Extractive–Abstractive Model for Legal Summarization,” International Journal of AdvancedComputerScience1and1Applications, vol.16, no.3,pp.98–108, 2025.Link: https://ijacsa.thesai.org/

[7] T. Y. S. S. Santosh, C. Jia, P. Goroncy, and M. Grabmair, “RELexED: Retrieval-Enhanced Legal Summarization with Exemplar Diversity,” Proceedings of the Legal NLP Workshop, pp. 1–10, 2025.Link: https://aclanthology.org/

[8] [M. S. Anbarasi, A. Mohammed, D. R., M. V., and S. Mohanraj, “Abstractive Summarization of

Indian Legal Documents Using T5 and QLoRA," IEEE Access, vol. 12, pp. 145210–145222, 2024.Link:https://ieeexplore.ieee.org/document/

[9] S. K. Nigam, T. Dubey, G. Sharma, N. Shallum, K. Ghosh, and A. Bhattacharya, "LegalSeg: Unlocking the Structure of Indian Legal Judgments Through Rhetorical Role Classification," Proceedings of EMNLP, pp. 3341–3352, 2024.Link: https://aclanthology.org/

[10] Artificial Intelligence and LawResearch Group, "Effectiveness in Retrieving Legal Precedents Using Cost-Efficient Language Models," Artificial Intelligence and Law, vol. 32, no. 4, pp. 489–506, 2024. Link: https://link.springer.com/journal/1050

[11] K. Swamy, O. Salgare, O. Sakhare, and S. Tamboli, "ValidEase: NLP-Based Simplification and Summarization of Legal Documents," International Journal of Legal Informatics, vol. 5, no. 2, pp. 65–78, 2024.Link: https://www.legaltechjournal.org/

[12] P. Bindal, V. Kumar, S. Rathore, and V. Bhatnagar, "AugAbEx: Advancing Extractive Case Summarization," ACM Transactions on Information Systems, vol. 42, no. 1, pp. 1–25, 2024.Link: https://dl.acm.org/

[13] Artificial Intelligence Review Research Team, "Legal Lay Summarization Using Large Language Models," Artificial Intelligence Review, vol. 55, no. 6, pp. 4571–4590, 2022.Link: https://link.springer.com/journal/10462

[14] Y. Altameemi, M. Altamimi, A. Alkhalil, D. Uliyan, and R. F. Mansour, "Argumentative Discourse Summarization Using BERT-Transformer Models," Expert Systems with Applications, vol. 198, p. 116742, 2022.Link: https://doi.org/10.1016/j.eswa.2022.116742

[15] Roy and A. Philip, "Legal Document Summarization Using Natural Language Processing," International Journal of Computer Applications, vol. 174, no. 28, pp. 1–6, 2022.Link: https://www.ijcaonline.org/

[16] S. Agarwal, "A Deep Learning Approach for Legal Document Summarization Using Pre-Trained Transformers," Journal of Artificial Intelligence Research, vol. 73, pp. 995–1021, 2022.Link: https://www.jair.org/

[17] M. Ghosh, "BERT-Based Abstractive Summarization for Legal Case Documents," Computational Linguistics Journal, vol. 47, no. 3, pp.601–620, 2021.Link: https://direct.mit.edu/coli

[18] Chalkidis, "Legal-BERT: The Muppets Straight Out of Law School," Findings of EMNLP, pp. 2898–2904, 2021.Link: https://aclanthology.org/2021.findings-emnlp.247/

[19] D. Bommarito and D. Katz, "GPT Models for Legal Text Understanding," Artificial Intelligence and Law, vol. 29, no. 2, pp. 123–135, 2021.Link: https://link.springer.com/journal/10506

[20] K. Yamada, "Automatic Legal Judgment Summarization Using Longformer," Proceedings of ACL, pp. 456–466, 2022.Link: https://aclanthology.org/

[21] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," Proceedings of ACL, pp. 7871–7880, 2020. Link:https://aclanthology.org/2020.acl-main.703/

[22] Raffel et al., "Exploring the Limits of Transfer Learning with T5," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.Link: https://www.jmlr.org/papers/v21/20-074.html

[23] Beltagy, M. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.Link: https://arxiv.org/abs/2004.05150

[24] V. Bhattacharya, "LegalSumm: A Dataset and Benchmark for Legal Document Summarization," Proceedings of ACL, pp. 203–213, 2020.Link: https://aclanthology.org/

[25] Katz, "BillSum: A Corpus for Automatic Summarization of US Legislation," Proceedings of ACL, pp. 889–894, 2019.Link: https://aclanthology.org/2019.acl-long.307/