IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

LEGAL DOCGEN USING AI: YOUR SMART DOC GENERATOR

[1]Atharv Patil, [2]Ayush Shah, [3]Kartik Bapna
[1]Student, [2]Student, [3]Student
[1]Department of Information Technology,
[1]MIT ADT University, Pune, India

Abstract : As generative Artificial Intelligence (AI) technologies become more potent tools for automating the labor-intensive process of legal document generation, the legal business is undergoing a profound transformation. This study investigates the use of generative AI models in the production of legal documents, including deep learning and natural language processing. It investigates the possible advantages, difficulties, and moral issues related to this novel strategy. The first section of this paper presents the state of artificial intelligence (AI) in the legal industry, emphasizing the notable developments in natural language generation models. The methodology part explores the technical aspects of teaching AI models to generate legal documents while considering the subtleties of different document kinds and the intricacies of legal language. Several case studies demonstrate how AI-powered legal document generators have been successfully implemented in actual situations, highlighting their ability to boost productivity and lower human error rates. The investigation also examines the critical component of guaranteeing the caliber and precision of AI-generated legal documents, illuminating the processes involved in assessment and confirmation. Concerns about data privacy, liability, and professional accountability are some of the ethical and legal issues regarding the use of AI in the legal field that are explored. The paper's conclusion highlights the changing nature of the legal profession by discussing potential future trends and ramifications of AI-driven legal document development. Considering this research, generative AI has the potential to change the legal sector by speeding document creation, saving time and money, and eventually improving access to legal services. This paper offers essential insights and recommendations for legal practitioners, legislators, and researchers as they navigate the changing legal landscape in the age of AI-powered document development. (Abstract)

IndexTerms **- Artificial Intelligence, tools, documentation, natural language generation, legal document**

I.INTRODUCTION

In the field of law, the painstaking craft of creating legal papers is a cornerstone of legal practice. From contracts and agreements to court pleadings and legal memoranda, creating such documents necessitates an uncompromising dedication to clarity, legal terminology, and rigorous adherence to established structures. However, the current environment of the legal profession is on the verge of a significant shift, brought in by rapid breakthroughs in artificial intelligence (AI), and notably generative AI.

Setting out on an ambitious trip, this research study investigates the complex dynamics and significant ramifications of "Legal Document Generation using Generative AI." Fundamentally, this is a paradigm change that combines the state-of-the-art powers of artificial intelligence (AI) with the centuries-old traditions of legal expertise. The main idea is simple: how can we use the amazing powers of generative AI, as demonstrated by models like GPT-3 and BERT, to enhance and innovate the process of creating legal documents?

In this situation, generative AI has enormous potential. It claims to simplify the difficult and frequently time-consuming process of drafting legal papers. Generative AI systems are positioned to significantly boost productivity, uphold remarkable uniformity, and significantly reduce the margin for human mistake by imitating human-like text production. All things considered, this technology has the potential to improve the accuracy and caliber of legal paperwork while also saving a great deal of time.

However, there are obstacles in the way of this evolution, and probably most importantly, these obstacles go beyond the domain of technology. Artificial intelligence (AI) in the legal profession always raises a number of problems about the ethical and legal implications of AI-generated content. One crucial point is still crucial as we enter this unexplored area: how can we make sure AI-generated documents are accurate, compliant with the law, and morally upright? These responses are crucial to the integrity of the legal system.

Hence, the goal of this research project is to thoroughly explore the complex field of "Legal Document Generation using Generative AI." Our objective is to provide an extensive overview of the historical development and current state of this revolutionary technique. Through a thorough analysis of the benefits, uses, and constraints of this cutting-edge technology, along with a focus on the critical role that human expertise plays in this intimate collaboration between legal practice and AI innovation, our goal is to provide light on this dynamic and complicated confluence.

As the legal industry ventures into this new technology frontier, it is critical to balance the benefits of greater efficiency with the demands of upholding the strictest legal and ethical guidelines. This research acts as a lighthouse, showing the way toward a time when legal knowledge and artificial intelligence will collaborate to completely transform the legal document generation industry.

II.Literature Survey.

In sequence modeling and transduction challenges like language modeling and machine translation, recurrent neural networks—long short-term memory and gated recurrent neural networks in particular—have solidly established themselves as state-of-the-art methods. Since then, many attempts have persisted in pushing the limits of encoder-decoder architectures and recurrent language models. Usually, recurrent models factor the calculation along the input and output sequences' symbol locations. They create a series of hidden states, ht, based on the input for position t and the preceding hidden state, ht−1, by aligning the locations with computation time steps. This sequential structure prevents training examples from being parallelized, which is important for longer sequence lengths when memory restrictions restrict batching among samples. Recent research has used conditional computation and factorization techniques to yield notable gains in computational efficiency and, in the case of the latter, improved model performance. However, sequential processing still has its basic limitations. With attention mechanisms, dependencies in input or output sequences may be modeled regardless of their distance from one another, making them an essential component of appealing sequence modeling and transduction models in a variety of activities. Nonetheless, such attention strategies are employed in all but a few instances. in conjunction with a recurrent network. The Transformer model architecture, which we present in this study, draws global interdependence between input and output by depending solely on an attention mechanism, rather than recurrence. After being trained for as little as twelve hours on eight P100 GPUs, the Transformer may achieve a new state of the art in translation quality and allows for substantially higher parallelization.

III.BACKGROUND

The Extended Neural GPU, ByteNet, and ConvS2S all use convolutional neural networks as their fundamental building blocks and compute hidden representations in parallel for all input and output points with the aim of minimizing sequential computation. For ConvS2S and ByteNet, the number of operations needed to relate signals from two arbitrary input or output positions increases linearly and logarithmically, respectively, with the distance between the positions. As a result, learning dependencies between far-off locations becomes more challenging. By averaging attention-weighted positions, the Transformer reduces this to a fixed number of operations; however, this comes at the expense of a lower effective resolution, which we mitigate with Multi-Head Attention, as we'll cover in section 3.2. In order to calculate a representation of a sequence, self-attention, also referred to as intra-attention, is an attention mechanism that links several points inside a single sequence. Numerous tasks, such as reading comprehension, abstractive summarization, textual entailment, and learning task-independent phrase representations, have shown success when self-attention is employed. End-to-end memory networks have demonstrated strong performance on language modeling and simple-language question answering tasks. These networks rely on a recurrent attention mechanism rather than sequence aligned recurrence. To the best of our knowledge, however, the Transformer is the first transduction model that does not use sequence aligned RNNs or convolution—rather, it solely relies on self-attention to calculate representations of its input and output. The Transformer will be explained, and self-attention will be encouraged, in the parts that follow.

IV.MODEL ARCHITECTURE

The Extended Neural GPU, ByteNet, and ConvS2S all use convolutional neural networks as their fundamental building blocks and compute hidden representations in parallel for all input and output points with the aim of minimizing sequential computation. For ConvS2S and ByteNet, the number of operations needed to relate signals from two arbitrary input or output positions increases linearly and logarithmically, respectively, with the distance between the positions. As a result, learning dependencies between far-off locations becomes more challenging. By averaging attention-weighted positions, the Transformer reduces this to a fixed number of operations; however, this comes at the expense of a lower effective resolution, which we mitigate with Multi-Head Attention (Fig 2), as we'll cover in section 3.2. To calculate a representation of a sequence, self-attention, also referred to as intra-attention, is an attention mechanism that links several points inside a single sequence. Numerous tasks, such as reading comprehension, abstractive summarization, textual entailment, and learning task-independent phrase representations, have shown success when self-attention is employed. End-to-end memory networks have demonstrated strong performance on language modeling and simple-language question answering tasks. These networks rely on a recurrent attention mechanism rather than sequence aligned recurrence. To the best of our knowledge, however, the Transformer is the first transduction model that does not use sequence aligned RNNs or convolution—rather, it solely relies on self-attention to calculate representations of its input and output. The Transformer will be explained, and self-attention will be encouraged, in the parts that follow.

Research Through Innovation

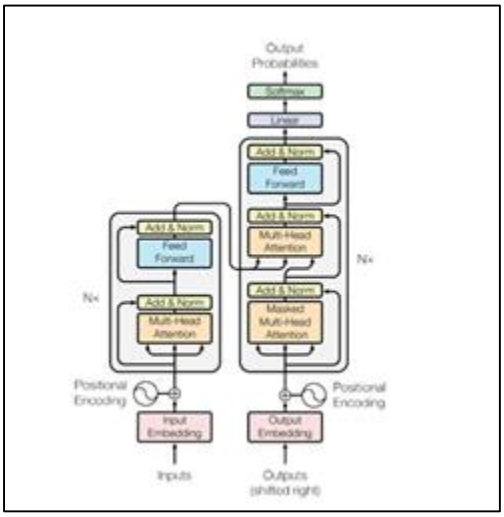

Figure 1: Transformer Architecture -by Ashish Vaswani

V.ENCODE AND DECODER STACKS

ENCODER

A stack of N = 6 identical layers makes up the encoder. A layer consists of two sublayers. A basic, position-wise fully connected feed-forward network is the second, while a multi-head self-attention mechanism is the first. For each of the two sub-layers, we use a residual connection, and then we apply layer normalization. In other words, each sub-layer's output is LayerNorm, and Sublayer is the function that each sub-layer implements on its own. All of the model's sub-layers and the embedding layers generate outputs with dimension d-model = 512 in order to support these residual connections.

DECODER

Additionally, a stack of N = 6 identical layers makes up the decoder. The decoder adds a third sub-layer, which handles multi-head attention over the encoder stack's output, in addition to the two sub-layers in each encoder layer. We use residual connections around each sub-layer, just like the encoder, and then layer normalization. Additionally, we alter the decoder stack's self-attention sub-layer to stop positions from paying attention to positions that come after. The forecasts for location i can only rely on the known outputs at positions less than i thanks to this masking and the output embeddings' one-position offset.

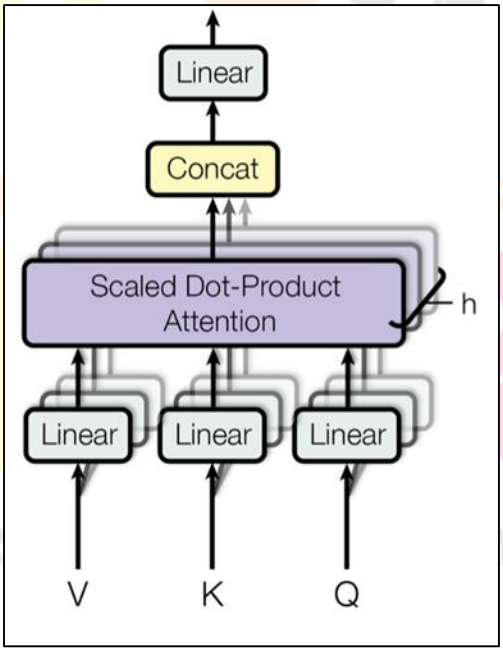

Figure 2: Multi head attention

ATTENTION FUNCTION

Mapping a query and a collection of key-value pairs to an output—where the query, keys, values, and output are all vectors—is how one might characterize an attention function. The results are calculated as a weighted sum of the values, with each value's weight determined by the query's compatibility function with its matching key.

TEXT-TO-TEXT TRANSFER TRANSFORMER

The Text-to-Text Transfer Transformer, or T5, is an architecture based on Transformers that employs a text-to-text methodology. All tasks are viewed as feeding the model text as input and training it to produce some goal text. These tasks include translation, question answering, and classification. This makes it possible to utilize the same model, loss function, hyperparameters, etc. for all the tasks that we have.

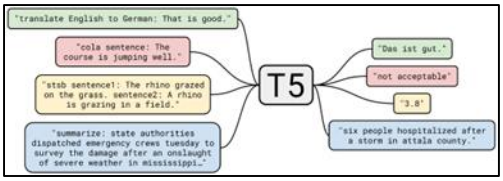

Figure 3: T5 (Text-To-Text Transfer) Transformer)

SUMMARY OF EXISTING TOOLS

Name of Available tools	Advantages	Limitations
LegalTemplates.net	Simple process for generating residential lease documents	Only available in United states.
Lawdepot.com	Various types of residential lease available.	Other kind of legal documents not available.
Indiafilings.com	Templates for various legal documents available.	Only templates are available.
Hellobonsai.com	Various types of contracts like freelance and client contracts available.	Contracts making is not the main goal of the website.
Amto	Use generative AI based on ChatGPT for drafting legal documents, letters, and emails.	Monthly fees are high, given word count and feature limitations and not recommended for complex or high value -legal documents
Detangle.ai	Get AI-generated summaries of lengthy documents, audio files, or video.	Need to join a new user waitlist before you can use the app and Per-file fees are expensive, even for shorter documents.
Lex Machina	Use Legal Analytics Quick Tools to compare judges, law firms, parties, and more.	Lex Machina isn't self-service—you'll need to work with their sales team to sign up and choose the tools you need.

APPLICATIONS OF ATTENTION IN OUR MODEL

The Transformer uses multi-head attention in three different ways:
In "encoder-decoder attention" layers, the encoder's output provides the memory keys and values, while the queries originate from the previous decoder layer.
This enables all points in the input sequence to be attended to by every position in the decoder. This reflects the common sequence-to-sequence models' encoder-decoder attention processes.
There are layers of self-attention in the encoder. All of the keys, values, and queries in a self-attention layer originate from the same source—in this example, the encoder's preceding layer's output. Every point in the encoder has the ability to service every position in the encoder's previous layer.
In a similar vein, the decoder's self-attention layers enable any position to attend to all positions up to and including that position. To maintain the auto-regressive feature, the decoder must stop leftward information flow. By masking out (setting to $-\infty$) all values in the SoftMax input that correspond to illegal connections, we implement this inside of scaled dot-product attention.

VI.CONCLUTION

In this study, we developed the Transformer, the first sequence transduction model based solely on attention, which uses multi-headed self-attention in place of the recurrent layers that are often seen in encoder-decoder designs. The Transformer may be trained much faster for translation tasks than systems built on convolutional or recurrent layers. The advantages are apparent: AI

revolutionizes the preparation of a wide range of legal documents, from contracts to legal studies, by providing speed, uniformity, and fewer errors. But there are difficulties with this evolution. Crucial factors to consider are ensuring legal and ethical compliance, minimizing prejudices, and upholding transparency. It is clear that human oversight and subject knowledge are still essential for preserving the integrity of the legal profession as we negotiate the junction of AI and legal practice. Although artificial intelligence (AI) enhances the ability of legal practitioners, the potential to transform the future of legal document generation lies in the synergistic combination of human judgment and AI-driven efficiency. The sector is dynamic, with continuous improvements and an increasing focus on moral and legal requirements, guaranteeing that future legal papers are effective and compliant with the highest moral and legal standards. In conclusion, the literature survey on "Legal Document Generator using Generative AI" reveals the transformative potential of artificial intelligence in the legal field. From historical developments to current trends, the integration of generative AI presents a promising avenue for enhancing efficiency, accuracy, and accessibility in legal document generation.

VII. ACKNOWLEDGMENT

We would like to thank Prof. Rahul Bhole, Assistant Professor, MIT ADT University, for their invaluable guidance and support throughout the research process. We also acknowledge the contributions of our authors Atharv Patil, Kartik Bapna and Ayush Shah for their technical assistance and support.

VIII. REFERENCES

[1] "Attention Is All You Need" by Ashish Vaswani (2017): This is the foundational paper for the Transformer architecture, which is the basis for many state-of-the-art generative models like GPT-3.
[2] "GPT-3" by Brown (2020): This paper introduces GPT-3, one of the largest and most powerful language models to date, which can be used for document generation tasks.
[3] "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin. (2018)
[4] "T5: Text-to-Text Transfer Transformer" by Raffel (2019): T5 is a text-to-text framework that can be adapted for various natural language generation tasks, including document generation.
[5] “Robust Deep Reinforcement Learning for Extractive Legal Summarization” by Duy-Hung Nguyen(2021) :This article provides insights to use reinforcement learning to train current deep summarization models to improve their performance in the legal domain
[6] "CTRL: A Conditional Transformer Language Model for Controllable Generation" by Keskar, Nitish S., et al. (2019)
[7] "XLNet: Generalized Autoregressive Pretraining for Language Understanding" Authors: Yang, Zhilin.: XLNet improves upon BERT by addressing its limitations, such as the inability to model bidirectional context. It introduces a novel permutation-based training approach that enables capturing bidirectional context while maintaining the advantages of autoregressive models.
[8] "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context" by Dai, Zihang: Published in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019) Transformer-XL extends the Transformer architecture to handle longer sequences by introducing a segment-level recurrence mechanism. This allows capturing dependencies beyond the fixed-length context window.
[9] "Self-Attention with Relative Position Representations" by: Shaw, Peter. Published in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018). This paper explores the use of relative position representations in self-attention mechanisms, which helps capture positional information efficiently and effectively.
[10] "Reformer: The Efficient Transformer" by Kitaev, Nikita Published in: arXiv preprint arXiv:2001.04451 (2020) Reformer proposes several optimizations to make Transformers more memory-efficient, enabling the training of larger models on longer sequences
[11] "A Structured Self-attentive Sentence Embedding" by Lin, Zhouhan. (2017) This paper introduces a structured self-attentive sentence embedding model, which learns to represent sentences by attending to different parts of the input based on their importance for the task at hand.
[12] BERT Rediscovers the Classical NLP Pipeline" by Tenney, Ian. (2019) This paper investigates how BERT, despite being trained end-to-end, implicitly learns to perform tasks that resemble traditional NLP pipeline components, such as part-of-speech tagging, named entity recognition, and dependency parsing.
[13] Longformer: The Long-Document Transformer" by Beltagy, Iz, (2020) Longformer extends the Transformer architecture to handle long documents by introducing a combination of local and global attention mechanisms, addressing the limitations of standard Transformers on tasks involving lengthy input sequences.
[14] "The Illustrated Transformer" by Jay Alammar. This online resource provides a detailed and visually intuitive explanation of the Transformer architecture, including its self-attention mechanism, layer normalization, and positional encoding.
[15] "BERTweet: A Pre-trained Language Model for English Tweets" by Tang, Raphael (2020)This paper introduces BERTweet, a variant of BERT pre-trained specifically for English tweets, highlighting the adaptability of Transformer-based architectures to different linguistic domains and data type.