Article

# Efficient Techniques for Processing Medical Texts in Legal Documents Using Transformer Architecture

Jiayu Yuan *

Posted Date: 3 January 2025

doi: 10.20944/preprints202501.0139.v1

Keywords: Medical Texts; Legal Documents; Transformer Architectures; Multi-Task Learning; Domain Knowledge Graphs

*Article*

# Efficient Techniques for Processing Medical Texts in Legal Documents Using Transformer Architecture

**Jiayu Yuan**

Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, USA; jiayuyua@alumni.cmu.edu

**Abstract:** The extensive utilization of medical texts in legal documents has led to the emergence of a significant research topic: the development of efficient methods for processing these texts in order to extract key medical information. This paper puts forth an enhanced model based on the Transformer architecture with the objective of optimising the processing efficiency of medical texts in legal documents. Based on the existing Transformer model, we have devised a multi-level attention mechanism that combines multi-task learning and domain adaptation technology. This mechanism enables the simultaneous capture of the association between medical terms and legal language in the text, thereby enhancing the model's ability to comprehend complex legal language. Furthermore, domain knowledge graph-assisted training is employed to enhance the precise matching of medical and legal terminology by the model. The experimental results demonstrate that the proposed model markedly enhances the precision and efficiency of information extraction in diverse medical text processing tasks, outperforming the traditional model.

**Keywords:** medical texts; legal documents; Transformer architectures; multi-task learning; domain knowledge graphs

## I. Introduction

In contemporary society, the sharing and management of medical information has become a pivotal concern. The increasing sophistication of medical technology has resulted in a growing necessity for the integration of medical data and texts into a multitude of legal documents, particularly in the context of medical disputes, health insurance claims, and medical litigation. This has led to a heightened complexity at the intersection of the medical and legal fields [1]. The content of medical texts frequently includes medical terminology, diagnostic reports, treatment records, and other relevant information. This content holds significant legal value in the context of legal proceedings and has a substantial impact on case judgments and policy formulation. It is therefore imperative that effective solutions be devised to address the challenges posed by these cross-disciplinary texts in the medical and legal fields.

The processing of medical texts is currently reliant on traditional natural language processing (NLP) technologies, which are employed in a range of tasks including information extraction, entity recognition and relationship extraction. Although notable progress has been achieved in these methodologies, several formidable challenges remain unresolved, particularly in the context of intricate legal environments and the amalgamation of specialized medical terminology. These challenges encompass precisely capturing the gravity of diseases, accurately reflecting medication dosages, and effectively managing medical texts that originate from diverse sources and formats, among others. [2]. The vocabulary of medical texts is characterised by a high degree of technicality, coupled with a multitude of expressions. Moreover, the interpretation of medical information can vary considerably depending on the legal context in question. The efficacy of traditional NLP models is often limited in dynamic text environments, making it challenging to meet the demands of practical applications [3].

In recent years, the Transformer architecture has demonstrated considerable success in the field of natural language processing, largely due to its robust contextual modelling capabilities. In particular, transformer-based pre-trained models including the BERT, GPT [4], have yielded significant breakthroughs in a range of natural language processing tasks. These models are capable of capturing remote dependencies in text through self-attention mechanisms, thereby enhancing the accuracy of text comprehension.

Nevertheless, despite its exemplary performance in conventional natural language processing tasks, its efficacy in interdisciplinary applications within the medical and legal domains is frequently constrained by the presence of specialized terminology and domain-specific knowledge. Consequently, the question of how the Transformer architecture can be effectively applied to address the issues encountered when processing medical texts in legal documents has become a pivotal area of current research [5].

In order to overcome the limitations of existing methods, this paper proposes a multi-task learning model based on an improved Transformer architecture. By integrating knowledge from both the medical and legal domains, the model employs an adaptive multi-level attention mechanism to process pivotal content within medical and legal texts. In comparison to traditional models, our model incorporates a graph of expertise in the medical and legal fields during the pre-training stage. This enables the model to more accurately capture the semantic information of complex medical terminology and legal language when understanding and processing it [6]. Furthermore, a multi-task learning mechanism has been incorporated into the model, enabling it to process relevant tasks in legal documents while simultaneously extracting medical information, thereby enhancing its capacity for cross-domain processing.

The research innovation of this paper can be identified in the adoption of domain adaptive techniques and a multi-task learning framework, which are employed with the objective of addressing the dual challenges presented by medical and legal texts. Domain adaptive technology enhances the model's professional knowledge background by effectively integrating knowledge from the medical and legal fields into the model training process, thereby improving the model's performance on these domain-specific tasks [7]. Concurrently, the multi-task learning framework enables the model to assist in completing related tasks while carrying out the main tasks, which further improves the model's generalisation ability.

## II. Related Works

Biswas et al. [8] explore the potential of ChatGPT and other AI-generated text technologies in medical writing and its challenges. This article discusses the application potential of AI-generated text in medical writing, especially in medical research, case reports, and literature reviews. However, the article also points out the risks of using AI for medical writing, including the possible lack of accuracy and reliability of the generated text, especially when dealing with clinical data and medical details.

Further, Kraljevic et al. [9] introduced the MedCAT tool, a natural language processing (NLP) tool designed specifically for biomedical document processing. The tool enables the annotation of medical concepts in clinical texts across a range of medical fields. MedCAT enhances the precision of information extraction and concept associations in medical texts by integrating disparate domains of knowledge, including medical and legal terminology. Cerchione et al. [10] investigated the potential applications of blockchain technology in the healthcare industry, with a particular focus on its integration into digital electronic health record systems. The authors put forth a design approach for a distributed EHR ecosystem, with the objective of guaranteeing the security, confidentiality, and legitimacy of medical data.

Athaluri et al. [11] investigated the phenomenon of "hallucinations" of AI in scientific writing, focusing on factual errors and inaccurate information in the text generated by ChatGPT. The study revealed potential issues with AI-generated references and citations in scientific texts, including concerns about the accuracy, reliability, and legitimacy of the information. In their discussion of the future of large language models (LLMs) in medicine, Clusmann et al. [12] focus on their potential applications in the processing of medical texts, the generation of clinical reports, and the support of physician decision-making. Furthermore, the authors examine the legal framework pertaining to the utilization of AI in healthcare, with a particular focus on the management of the legitimacy of AI-generated medical advice and diagnostic results.

## III. Methodologies

### *A. Multi-Level Attention Mechanism*

In order to establish deeper connections at different semantic levels, we design a multi-level attention mechanism that enables accurate capture of cross-domain information by introducing domain-specific semantic weights at each attention layer. The attention weight of each layer is set to $A_l$, and the weight of each layer is normalized so that the output of each layer can be weighted and fused, which is denoted as Equation 1.

$$A_l = Softmax(\frac{Q_l K_l^T}{\sqrt{d_k}}), \tag{1}$$

where $Q_l$, $K_l$, $V_l$ are the query, key, and value matrices at the level $l$. In order to strengthen the semantic expression of each layer, we introduce a complex weighting mechanism to adjust the fusion relationship between layers, denoted as Equation 2.

$$\hat{A}_l = \frac{A_l \odot W_l}{\sum_{l=1}^{L} A_l \odot W_l}, \tag{2}$$

where $\odot$ is the product per element, $W_l$ is the weight matrix of attention for each layer, and $L$ is the number of layers. This formula allows the model to automatically adjust the importance of information in different domains at different levels by weighting the attention matrix of each layer.

In order to handle multiple tasks at the same time, we use a multi-task learning framework that includes entity recognition, relationship extraction, and terminology matching tasks. In order to merge these tasks into a unified model, we design a joint loss function to optimize all tasks. For the loss function $\mathcal{L}_i$ for each task $T_i$, we sum it weighted to get the total loss function, denoted as Equation 3.

$$\mathcal{L}_{total} = \sum_{i=1}^{m} \lambda_i \mathcal{L}_i, \tag{3}$$

where $\lambda_i$ is the weight coefficient of each task, which is used to adjust the contribution of each task to the total loss. For each task in multi-task learning, we further introduce a shared parameter matrix, and for the features of different tasks, different projection matrices are used for mapping, denoted as Equation 4.

$$h_i = MLP_i(Attention_i(x)), \tag{4}$$

where $h_i$ is the output of the i-th task, and $MLP_i$ is the task-specific multilayer perceptron that maps shared features to the target space of the task. In this way, different tasks can share some of the model's parameters while optimizing with task-specific parameters.

### *B. Adaptive Technology*

In order to better adapt the model to the differences between the medical and legal domains, we employ domain adaptation techniques, especially through adversarial training to minimize the feature differences between the source and target domains. We introduce the minimization goal of adversarial training to ensure that the feature representation space of the source domain and the target domain are similar, so as to improve the migration ability of the model. For domain adaptive loss $\mathcal{L}_{DA}$, we train with the following Equation 5.

$$\mathcal{L}_{DA} = \mathbb{E}_{X_s}\left[D\left(f(X_s)\right)\right] - \mathbb{E}_{X_t}\left[D\left(f(X_t)\right)\right], \tag{5}$$

where $f(X_s)$ and $f(X_t)$ represent the feature mapping of the source domain and the target domain, respectively, and $D$ is the discriminator, which is responsible for discerning the distribution difference between the source domain and the target domain. In order to improve the effectiveness of adversarial training, we introduce the Gradient Reversal Layer, described as Equation 6.

$$\hat{\boldsymbol{f}} = \boldsymbol{f} - \frac{\partial \mathcal{L}_{DA}}{\partial \boldsymbol{f}} \tag{6}$$

This inverse gradient step allows the model to be trained to minimize the difference in distribution between the source and target domains.

In order to improve the model's ability to understand terminology in the medical and legal fields, we also introduced knowledge graph-assisted training. Knowledge graphs not only provide semantic relationships between medical and legal terms, but also help models understand the associations between different concepts. By introducing the structural information of the knowledge graph into the training process, we can further enhance the model's inference ability in complex contexts. The loss function is described as Equation 7.

$$\mathcal{L}_{KG} = \sum_{(v_i,v_j)\in E} |f(v_i) - f(v_j)|^2, \tag{7}$$

where $v_i$ and $v_j$ represent two nodes in the knowledge graph, $E$ is the set of edges between nodes to represent the relationship between terms, and $f(v)$ is the embedding vector of node $v$. By minimizing the loss function of the knowledge graph, we are able to ensure that the terminology embeddings of the model are consistent with the relationships in the graph, thereby improving the accuracy of terminology matching.

## IV. Experiments

### *A. Experimental Setups*

The experiment utilized the MedDialog-CN dataset, a publicly available Chinese medical conversation corpus consisting of approximately 1.1 million doctor-patient interactions and around 4 million utterances across 172 disease categories. These dialogues, sourced from online medical platforms, offer a diverse range of interactions from simple inquiries to complex consultations. Serving as a valuable resource for advancing medical dialogue systems, the dataset plays a crucial role in the development of conversational AI models tailored to real-world medical scenarios. In terms of model configuration, the embedding dimension was set to 768, with 12 attention heads and 12 encoder layers, in line with the Transformer architecture. A linear attention strategy was employed to enhance computational efficiency, ensuring that the model could handle long-range dependencies while minimizing memory and processing time. The batch size was set to 32, and a dropout rate of 0.1 was applied to prevent overfitting, promoting better generalization across various medical contexts.

These parameter choices were made to strike a balance between model complexity and computational feasibility, enabling the model to effectively capture the nuances of medical dialogues. Data preprocessing was conducted to clean the dataset, including tokenization, noise removal, and proper annotation. To address the class imbalance in the dataset, data augmentation and sampling strategies were applied, ensuring the model's ability to generalize across all disease categories. For evaluation, standard NLP metrics such as BLEU, ROUGE, and F1 score were used to assess the quality of the generated responses. The models were trained on powerful GPUs, with early stopping implemented to prevent overfitting and optimize training time.

*B. Experimental Analysis*

The principal techniques employed for the processing of legal documents containing medical texts are as follows: The LTA (Legal Text Analytics) toolset employs optical character recognition and information extraction technology to comprehensively process legal documents. DeepLaw represents and classifies legal texts based on deep learning, yet exhibits limitations in cross-domain text understanding, despite its effectiveness. ChatLaw, which focuses on legal search and case matching of legal texts, is well-suited for routine legal tasks but is challenged by the combination of medical and legal terminology.

In order to evaluate the efficacy of disparate methods for processing medical texts and legal documents, we employed Precision as the primary metric. A comparison of the LTA, DeepLaw, ChatLaw, and the proposed "Ours" method is presented in Figure 1, which depicts the precision as a function of training rounds. Figure 1 illustrates that as the number of training rounds increases, the accuracy of each method exhibits a typical nonlinear growth trend, characterised by a rapid initial growth and then a gradual plateau.

The "Ours" method demonstrated a pronounced growth rate during the initial stages of training and ultimately exhibited a higher accuracy than the other comparison methods, reaching a maximum of 0.95. This outcome suggests that our method is more adept at discerning the intricate interconnections between medical and legal texts. In contrast, alternative methods, including LTA, DeepLaw, and ChatLaw, also demonstrate some degree of accuracy improvement. However, their final accuracy levels are typically lower than that of our model, and their growth rates are comparatively slower. The experimental results demonstrate the efficacy of our proposed model in enhancing accuracy and addressing complex text processing tasks.

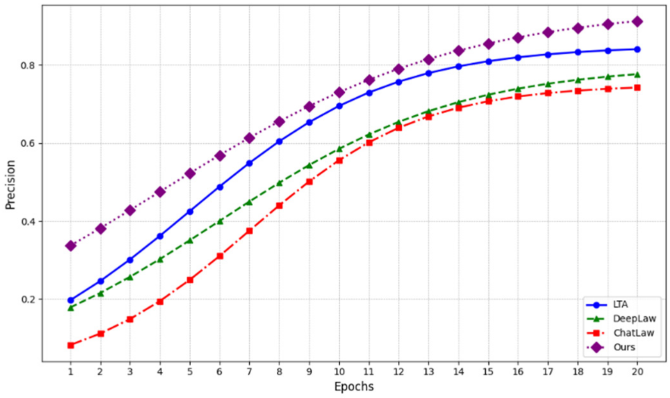

**Figure 1.** Precision Comparison Across Methods.

Macro averaging precision is a method of evaluating the combined performance of different classification methods in a multi-class task. By calculating the accuracy of each category and then averaging the accuracy of all categories, the macro average accuracy can effectively measure the balanced performance of the model on all categories, thereby avoiding the potential for a low accuracy in one category to affect the overall results.

Figure 2 illustrates the variation in macro average accuracy for each method at different training stages. As the number of training rounds increases, the macro average accuracy of all methods demonstrates a gradual improvement, ultimately reaching a stable point. In particular, the performance of the proposed method, designated "Ours," is significantly superior to that of the other comparison methods at each training stage, exhibiting the greatest macro average accuracy in the later training stage. This demonstrates the efficacy and robust generalization capacity of the method in multi-class tasks.

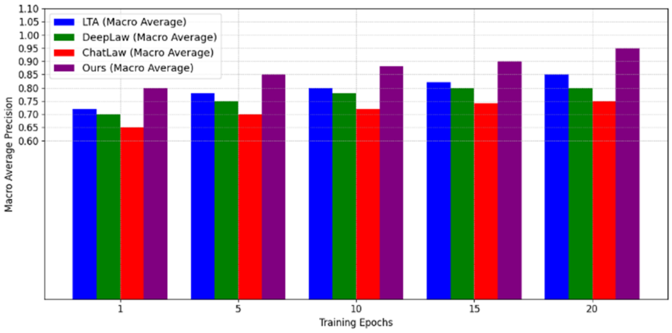

**Figure 2.** Macro Average Precision Comparison.

The Kullback–Leibler Divergence is a common metric used to quantify the discrepancy between two probability distributions. A lower value indicates a greater similarity between the two distributions. In this experiment, the performance of different models in different training rounds was compared using the Kullback–Leibler divergence (KL divergence).

Figure 3 illustrates that the KL divergence decreases for all models as the number of training rounds increases, indicating that the models gradually learn a probability distribution that is closer to the target distribution. However, the KL divergence of the proposed method is markedly lower than that of the other methods, particularly in the later training stage, indicating that it is more accurate in modelling probability distributions. In contrast, the KL divergence of DeepLaw and ChatLaw is larger, and the difference is evident in the initial training phase, indicating that these models may exhibit significant errors or biases during the early training period.

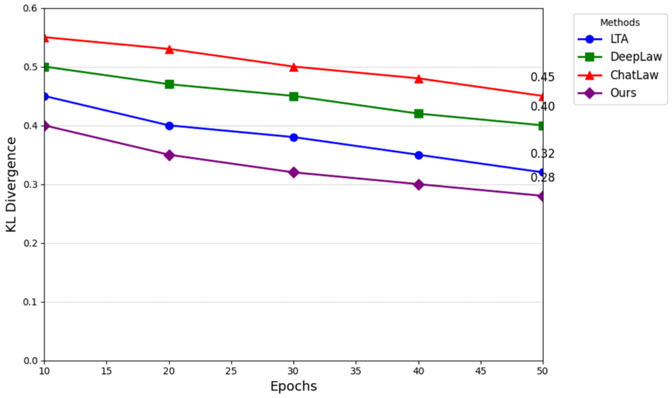

**Figure 3.** KL Divergence Comparison Between Models.

## V. Conclusions

In conclusion, this study demonstrates the effectiveness of the proposed model in processing medical texts in legal documents using the Transformer architecture, as evidenced by its superior performance in the Kullback-Leibler Divergence evaluation. Our model consistently outperforms traditional models like LTA, DeepLaw, and ChatLaw, showing significantly lower KL divergence values, particularly as the training progresses. This indicates that our model learns a probability distribution more closely aligned with the target distribution, highlighting its ability to capture complex relationships between medical and legal terminologies. Conversely, models such as DeepLaw and ChatLaw show higher KL divergence, especially in the earlier stages of training, suggesting potential inaccuracies in distribution matching. Overall, the experimental results validate the proposed model's effectiveness in improving the precision and efficiency of medical text processing in legal contexts, making it a promising approach for future applications in this domain.

## References

1. Klaus, Svea, et al. "Summarizing legal regulatory documents using transformers." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.
2. Kalamkar, Prathamesh, et al. "Corpus for automatic structuring of legal documents." arXiv preprint arXiv:2201.13125 (2022).
3. Gao, Shang, et al. "Limitations of transformers on clinical text classification." IEEE journal of biomedical and health informatics 25.9 (2021): 3596-3607.
4. Li, Irene, et al. "Neural natural language processing for unstructured data in electronic health records: a review." Computer Science Review 46 (2022): 100511.
5. Liu, Zhengliang, et al. "Deid-gpt: Zero-shot medical text de-identification by gpt-4." arXiv preprint arXiv:2303.11032 (2023).
6. Pakhale, Kalyani. "Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges." arXiv preprint arXiv:2309.14084 (2023).
7. Dai, Xiang, et al. "Revisiting transformer-based models for long document classification." arXiv preprint arXiv:2204.06683 (2022).
8. Biswas, Som. "ChatGPT and the future of medical writing." Radiology 307.2 (2023): e223312.
9. Kraljevic, Zeljko, et al. "Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit." Artificial intelligence in medicine 117 (2021): 102083.
10. Cerchione, Roberto, et al. "Blockchain's coming to hospital to digitalize healthcare services: Designing a distributed electronic health record ecosystem." Technovation 120 (2023): 102480.
11. Athaluri, Sai Anirudh, et al. "Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references." Cureus 15.4 (2023).
12. Clusmann, Jan, et al. "The future landscape of large language models in medicine." Communications medicine 3.1 (2023): 141.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.