IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 October 2024, accepted 9 November 2024, date of publication 12 November 2024,
date of current version 20 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3496666

RESEARCH ARTICLE

LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration

XURAN WANG[1], XINGUANG ZHANG[2], (Member, IEEE), VANESSA HOO[3], ZHOUHANG SHAO[4], AND XUGUANG ZHANG[5], (Associate Member, IEEE)

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080, USA
[3]School of Mathematics, School of Economics, Georgia Institute of Technology, Atlanta, GA 30332, USA
[4]Department of Computer Science and Engineering, The University of California, San Diego, La Jolla, CA 92093, USA
[5]School of Business, Computing and Social Sciences, University of Gloucestershire, GL50 2RH Cheltenham, U.K.

Corresponding author: Xuran Wang (xurwang@seas.upenn.edu)

ABSTRACT Legal judgment prediction (LJP) presents a formidable challenge in artificial intelligence, demanding intricate comprehension of legal texts, nuanced interpretation of statutes, and complex reasoning over multifaceted case elements. While recent advancements in natural language processing have shown promise, existing approaches often struggle to capture the sophisticated interplay between facts, legal principles, and precedents that characterize legal decision-making. This paper introduces LegalReasoner, a novel multi-stage framework that leverages large language models (LLMs) and integrates domain-specific knowledge for enhanced legal judgment prediction. Our approach encompasses four key stages: 1) legal knowledge infusion, where we pre-train an LLM on a vast corpus of legal literature using contrastive learning techniques; 2) case-law retrieval, employing a graph neural network to identify relevant precedents and statutes; 3) multi-hop reasoning, utilizing a transformer-based architecture with a hierarchical attention mechanism to navigate complex legal arguments; and 4) judgment synthesis, where we employ a generative adversarial network to produce coherent and legally sound judgments. We evaluate LegalReasoner on two diverse datasets: the European Court of Human Rights (ECHR) cases and the Chinese AI and Law Challenge (CAIL2018). Our framework demonstrates substantial improvements over state-of-the-art baselines, achieving an average accuracy increase of 7.8% across all datasets. Furthermore, we conduct extensive ablation studies and interpretability analyses to elucidate the contributions of each component and provide insights into the model's decision-making process. Our work not only advances the field of automated legal reasoning but also offers a transparent and explainable system that could serve as a valuable tool for legal professionals. By bridging the gap between AI and legal expertise, LegalReasoner paves the way for more efficient, consistent, and fair legal decision-making processes.

INDEX TERMS Legal judgment prediction, large language models, knowledge integration, multi-hop reasoning.

I. INTRODUCTION

The field of legal judgment prediction (LJP) stands at the intersection of artificial intelligence and jurisprudence, presenting a formidable challenge that demands the synthesis of advanced computational techniques with intricate legal knowledge. As judicial systems worldwide grapple with increasing caseloads and the complexities of modern law, the development of accurate and reliable LJP systems has become not just an academic pursuit, but a pressing societal need [1]. LJP systems aim to forecast the outcomes of legal cases based on the facts presented, relevant statutes, and historical precedents.

However, the task is fraught with challenges that push the boundaries of current artificial intelligence capabilities. First, legal texts are characterized by domain-specific terminology,

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

intricate sentence structures, and nuanced meanings that often elude standard natural language processing techniques [2]. Second, legal decision-making involves complex reasoning processes that consider multiple factors simultaneously, including factual evidence, legal principles, precedents, and sometimes conflicting interpretations of the law [3]. Third, the law is not static; it evolves over time through new legislation and landmark cases. An effective LJP system must account for these temporal dynamics and adapt its predictions accordingly [4]. Finally, in the legal domain, the reasoning behind a prediction is often as important as the prediction itself. LJP systems must provide transparent and interpretable decisions to be truly useful in legal practice [5].

Recent advancements in natural language processing, particularly the development of large language models (LLMs) like GPT-3 [6], have shown promise in tackling some of these challenges. LLMs demonstrate remarkable capabilities in understanding and generating human-like text, even in specialized domains. However, their application to LJP has been limited, primarily due to the lack of domain-specific legal knowledge and the need for structured reasoning that aligns with legal principles.

To address these limitations and advance the state-of-the-art in LJP, we introduce LegalReasoner, a novel multi-stage framework that leverages the power of LLMs while integrating domain-specific legal knowledge and structured reasoning processes. Our approach consists of four key stages: First, we pre-train an LLM on a vast corpus of legal literature using contrastive learning techniques. This allows the model to acquire deep domain-specific knowledge and understand the nuances of legal language. Second, employing a graph neural network, we identify relevant precedents and statutes for each case. This stage mimics the legal research process, providing a foundation for informed decision-making. Then, utilizing a transformer-based architecture with a hierarchical attention mechanism, we enable the model to navigate complex legal arguments, considering multiple factors and their interactions. Finally, We employ a generative adversarial network to produce coherent and legally sound judgments, ensuring that the output not only predicts the outcome but also provides a well-reasoned explanation.

We evaluate LegalReasoner on two diverse datasets: the European Court of Human Rights (ECHR) cases [2] and the Chinese AI and Law Challenge (CAIL2018) [7]. This comprehensive evaluation across different legal systems and languages demonstrates the robustness and versatility of our approach.

The main contributions of this paper are as follows:

1) We propose LegalReasoner, a novel multi-stage framework that integrates LLMs, domain-specific knowledge, and structured reasoning for enhanced legal judgment prediction.
2) We introduce innovative components, including contrastive learning for legal knowledge infusion, graph-based case-law retrieval, and a hierarchical multi-hop reasoning mechanism, which together significantly advance the state-of-the-art in LJP.
3) We conduct extensive experiments on two diverse, large-scale datasets, demonstrating substantial improvements over existing baselines with an average accuracy increase of 7.8%.
4) We provide in-depth ablation studies and interpretability analyses, offering valuable insights into the model's decision-making process and the contribution of each component.

Our work not only pushes the boundaries of AI in legal applications but also offers a transparent and explainable system that could serve as a valuable tool for legal professionals. By bridging the gap between AI capabilities and legal expertise, LegalReasoner paves the way for more efficient, consistent, and fair legal decision-making processes.

The rest of this paper is organized as follows: Section II reviews related work in legal judgment prediction, large language models, and knowledge integration in AI. Section III introduces the problem formulation and foundation of other techniques in this work. Section IV presents the detailed methodology of our LegalReasoner framework. Section V describes the experimental setup, datasets, and results. Also, the corresponding in-depth analyses, ablation studies, and case studies are provided in this section. Finally, Section VI concludes the paper and discusses future research directions and potential societal impacts of our work.

II. RELATED WORKS

Our work builds upon and extends several key areas of research in artificial intelligence and law. In this section, we review relevant literature in legal judgment prediction, large language models, and knowledge integration in AI.

A. LEGAL JUDGMENT PREDICTION

Legal Judgment Prediction (LJP) has been an active area of research in the intersection of AI and law. Early approaches relied on manual feature engineering and traditional machine learning algorithms. For instance, [8] employed support vector machines with hand-crafted features to predict the outcomes of Chinese criminal cases. With the advent of deep learning, more sophisticated models have been proposed. Reference [9] introduced a hierarchical attention network for LJP, which could capture the hierarchical structure of legal documents. Reference [10] proposed TOPJUDGE, an iterative model that predicts the judgment by extracting relevant paragraphs from similar cases. More recently, transformer-based models have shown promising results in LJP. Reference [2] employed a hierarchical version of BERT to encode both the facts and relevant articles for judgment prediction. References [11] and [12] revealed that the legal-BERT model [13] pre-trained on a large corpus of legal documents, demonstrated improved performance on various legal NLP tasks, including LJP. Despite these advancements, existing approaches often struggle to capture the complex

reasoning process involved in legal decision-making and lack the ability to generate explanatory judgments.

B. LARGE LANGUAGE MODELS IN LEGAL DOMAIN

Large Language Models (LLMs) have recently gained attention in the legal domain due to their impressive performance in various natural language processing tasks. Zhong et al. [1] provided a comprehensive survey of how NLP techniques, including LLMs, benefit legal systems. Chalkidis et al. [13] introduced legal-BERT, a BERT model pre-trained on legal corpora, which outperformed generic BERT on several legal NLP tasks. However, these models are primarily used for understanding legal texts rather than generating judgments or explicating legal reasoning. Brown et al. [14] demonstrated the fine-tuning capabilities of GPT-3 on various tasks, including some simple legal reasoning and legal rule classification tasks. However, the application of such large models to complex LJP tasks remains largely unexplored.

C. KNOWLEDGE INTEGRATION IN AI SYSTEMS

Integrating domain-specific knowledge into AI systems has been a long-standing challenge. In the context of LLMs, [15] proposed knowledge-enhanced language models that incorporate external knowledge bases during pre-training. Reference [16] introduced the retrieval-augmented generation model, which combines a neural retriever with a sequence-to-sequence model for knowledge-intensive tasks. In the legal domain, [1] explored how to leverage external legal knowledge to improve the performance and interpretability of LJP models. Our work extends these ideas by proposing a multi-stage framework that integrates legal knowledge at different levels of abstraction, from pre-training to reasoning and generation.

By building upon and integrating these diverse areas of research, our work aims to advance the state-of-the-art in legal judgment prediction, offering a comprehensive framework that addresses the unique challenges of this complex task.

III. PRELIMINARIES

Before delving into the details of our LegalReasoner framework, we first formalize the legal judgment prediction task and introduce key concepts and techniques that form the foundation of our approach.

A. PROBLEM FORMULATION

Legal Judgment Prediction (LJP) can be formalized as a sequence-to-sequence learning problem with additional structured inputs. Given a case c, consisting of a factual description x, a set of relevant laws $L = \{l_1, l_2, \ldots, l_n\}$, and historical precedents $P = \{p_1, p_2, \ldots, p_m\}$, the goal is to predict the judgment y. Formally, we aim to learn a function f that maps the input to the output:

$$f : (x, L, P) \rightarrow y \tag{1}$$

where y can be a multi-dimensional output including the decision (e.g., guilty/not guilty), relevant articles of law, and the reasoning behind the decision. The challenge lies in capturing the complex interactions between the case facts, applicable laws, and relevant precedents, while also producing a coherent and legally sound judgment.

B. LARGE LANGUAGE MODELS

Large Language Models (LLMs) form a critical component of our framework. These models, typically based on the Transformer architecture [17], are pre-trained on vast corpora of text and have demonstrated remarkable capabilities in understanding and generating human-like text. Formally, an LLM can be viewed as a conditional probability distribution:

$$p(x_t | x_{<t}, \theta) \tag{2}$$

where x_t is the token at position t, $x_{<t}$ represents all previous tokens, and θ are the model parameters. In the context of LJP, we leverage LLMs for their ability to understand complex legal language and generate coherent judgments. However, we enhance their capabilities through domain-specific fine-tuning and structured reasoning mechanisms.

C. CONTRASTIVE LEARNING

Contrastive Learning is a technique we employ to infuse domain-specific legal knowledge into our model. The core idea is to learn representations that bring semantically similar inputs closer in the embedding space while pushing dissimilar inputs apart. Given an anchor sample x_a, a positive sample x_p (semantically similar to x_a), and a set of negative samples $\{x_n\}$, the contrastive loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(s(h_a, h_p)/\tau)}{\exp(s(h_a, h_p)/\tau) + \sum_n \exp(s(h_a, h_n)/\tau)} \tag{3}$$

where $s(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity), h_a, h_p, and h_n are the representations of the anchor, positive, and negative samples respectively, and τ is a temperature parameter.

D. GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs) are employed in our case-law retrieval stage to model the complex relationships between legal cases, statutes, and concepts. A GNN operates on a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. The key operation in a GNN is the message passing function, which updates node representations based on their neighbors:

$$h_v^{(l+1)} = \text{UPDATE}(h_v^{(l)}, \text{AGGREGATE}(\{h_u^{(l)} : u \in \mathcal{N}(v)\})) \tag{4}$$

where $h_v^{(l)}$ is the representation of node v at layer l, $\mathcal{N}(v)$ is the set of neighbors of v, and UPDATE and AGGREGATE are learnable functions.

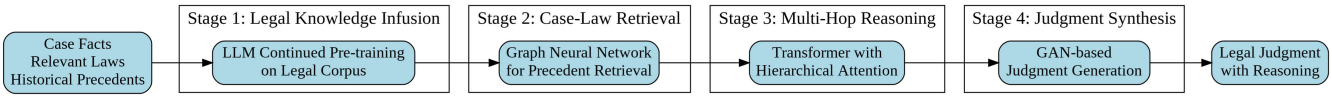


FIGURE 1. Overview of the LegalReasoner framework.

E. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) are used in our judgment synthesis stage to generate coherent and legally sound judgments. A GAN consists of two components: a generator G and a discriminator D. The generator aims to produce realistic judgments from an input noise vector z and case representation c:

$$\hat{y} = G(z, c) \tag{5}$$

The discriminator tries to distinguish between real judgments y and generated judgments $\hat{y}$:

$$D(y) \to [0, 1], \quad D(\hat{y}) \to [0, 1] \tag{6}$$

The two components are trained adversarially, with the generator trying to fool the discriminator and the discriminator trying to correctly classify real and generated judgments.

These preliminaries provide the necessary background for understanding the components of our LegalReasoner framework. In the following sections, we will elaborate on how these concepts are integrated and extended to address the unique challenges of legal judgment prediction.

IV. METHODOLOGY

In this section, we present a detailed description of our LegalReasoner framework. The architecture is designed to address the unique challenges of legal judgment prediction (LJP) through a multi-stage approach that combines the power of large language models with domain-specific knowledge and structured reasoning. Figure 1 provides an overview of the LegalReasoner framework.

A. LEGAL KNOWLEDGE INFUSION

The first stage of our framework aims to imbue the large language model with comprehensive legal knowledge. This is crucial because standard LLMs, while powerful in general language understanding, often lack the specialized knowledge required for legal reasoning. However, legal texts are replete with domain-specific terminology, complex concepts, and intricate relationships between legal principles. A model that can effectively predict legal judgments must first have a deep understanding of this specialized language and knowledge base.

Thus, we employ a contrastive learning technique to pre-train our LLM on a vast corpus of legal literature. The corpus includes legal textbooks, case law, statutes, and academic articles, covering multiple jurisdictions and areas of law. The contrastive learning objective is defined as:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(s(h_i, h_i^+)/\tau)}{\sum_{j=1}^{N} \exp(s(h_i, h_j)/\tau)} \tag{7}$$

where $s(\cdot, \cdot)$ is the cosine similarity, h_i is the representation of an anchor sample, h_i^+ is a positive sample (e.g., a paraphrase or related legal concept), h_j are negative samples, τ is a temperature parameter, and N is the number of samples in a mini-batch. This approach encourages the model to learn representations that cluster similar legal concepts together while pushing dissimilar concepts apart in the embedding space.

B. CASE-LAW RETRIEVAL

The second stage focuses on identifying relevant precedents and statutes for a given case. This mimics the legal research process and provides a foundation for informed decision-making. In fact, legal reasoning often relies heavily on precedent and statutory interpretation. By identifying and incorporating relevant prior cases and applicable laws, we enable the model to make predictions that are grounded in established legal principles.

Specifically, we employ a graph neural network (GNN) to model the complex relationships between cases, statutes, and legal concepts. The graph $G = (V, E)$ is constructed where nodes V represent cases and statutes, and edges E represent citations, shared legal concepts, or similar fact patterns. The GNN update rule for node representations is:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{AGGREGATE}\left(\{h_u^{(l)} : u \in \mathcal{N}(v)\}\right) + b^{(l)}\right) \tag{8}$$

where $h_v^{(l)}$ is the representation of node v at layer l, $\mathcal{N}(v)$ is the set of neighbors of v, $W^{(l)}$ and $b^{(l)}$ are learnable parameters, and σ is a non-linear activation function. For a given input case, we use the final node representations to retrieve the K most relevant precedents and statutes based on cosine similarity.

C. MULTI-HOP REASONING

The third stage of LegalReasoner involves a structured reasoning process that navigates the complex landscape of legal arguments. Legal decision-making often requires considering multiple, interrelated factors and following chains of reasoning. A simple classification approach fails to capture this nuanced process. Our multi-hop reasoning module aims to emulate the step-by-step logical progression that characterizes legal analysis.

We implement a transformer-based architecture with a hierarchical attention mechanism. The input to this module is the original case text, augmented with the relevant precedents and statutes retrieved in the previous stage. The multi-hop reasoning process is defined as:

$$r_i = \text{Transformer}([\text{CLS}; x; p_1; \ldots; p_K; s_1; \ldots; s_M]) \quad (9)$$

where x is the input case text, $p_1, \ldots, p_K$ are the retrieved precedents, $s_1, \ldots, s_M$ are the relevant statutes, and [CLS] is a special token for classification. The hierarchical attention mechanism allows the model to focus on different levels of information:

$$\alpha_i = \text{softmax}(w^T \tanh(Wr_i + b)) \quad (10)$$

$$o = \sum_i \alpha_i r_i \quad (11)$$

where w, W, and b are learnable parameters, and o is the final output representation.

D. JUDGMENT SYNTHESIS

The final stage of our framework aims to generate a coherent and legally sound judgment based on the reasoning process from the previous stages. In legal practice, the reasoning behind a decision is often as important as the decision itself. Therefore, our model must not only predict the outcome but also provide a well-reasoned explanation that adheres to legal writing conventions.

We utilize a generative adversarial network (GAN) for judgment synthesis. The generator G aims to produce realistic judgments, while the discriminator D learns to distinguish between real and generated judgments. The generator is trained with a combination of adversarial loss and a legal coherence loss:

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[\log D(G(z, o))] + \lambda \mathcal{L}_{\text{coherence}} \quad (12)$$

where z is random noise, o is the output from the multi-hop reasoning stage, and λ is a weighting parameter. The legal coherence loss $\mathcal{L}_{\text{coherence}}$ ensures that the generated judgment cites relevant precedents and statutes, maintains logical consistency, and follows the structure of legal arguments. The discriminator is trained to maximize:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z, o)))] \quad (13)$$

where p_{data} is the distribution of real judgments.

E. TRAINING AND OPTIMIZATION

The entire LegalReasoner framework is trained end-to-end using a multi-task learning objective that combines the losses from each stage:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{CL}} + \beta \mathcal{L}_{\text{GNN}} + \gamma \mathcal{L}_{\text{MHR}} + \delta(\mathcal{L}_G + \mathcal{L}_D) \quad (14)$$

where α, β, γ, and δ are weighting parameters that balance the contribution of each component. We use the Adam optimizer with a learning rate of 1e-5 and implement a warm-up strategy followed by linear decay. To handle the computational complexity, we employ gradient accumulation and mixed-precision training. This multi-stage approach allows LegalReasoner to capture the nuances of legal reasoning while leveraging the power of large language models and structured knowledge integration.

V. EXPERIMENTS

In this section, we present a comprehensive evaluation of our LegalReasoner framework. We first describe the datasets and baselines used in our experiments, followed by detailed experimental setup and evaluation metrics. We then present and analyze the results, including performance comparisons, ablation studies, and case studies to demonstrate the effectiveness and interpretability of our approach.

A. DATASETS

We evaluate our framework on two diverse and challenging legal datasets:

- **ECHR Cases** [2]: This dataset consists of 11,478 cases from the European Court of Human Rights (ECHR). Each case includes a factual description and a binary label indicating whether there was a violation of a specific article of the European Convention on Human Rights. We use the pre-defined split of 9,000 cases for training, 1,000 for validation, and 1,478 for testing.
- **CAIL2018** [7]: This dataset is from the Chinese AI and Law challenge, containing 2,676,075 criminal cases from the Supreme People's Court of China. Each case includes fact descriptions, applicable laws, charges, and prison terms. To manage computational resources, we randomly sample 200,000 cases for training, 20,000 for validation, and 20,000 for testing.

B. BASELINES

We compare LegalReasoner with the following state-of-the-art baselines:

- **TFIDF+SVM** [18]: A traditional machine learning approach using TF-IDF features and a Support Vector Machine classifier.
- **LSTM** [19]: A Long Short-Term Memory network that processes the input text sequentially, designed for legal judgment prediction task.
- **CNN** [4]: A Convolutional Neural Network that captures local features in the input text.
- **HIER-BERT** [2]: A BERT-based model that encodes the facts and articles using a hierarchical BERT architecture.
- **LegalBERT** [13]: A BERT model pre-trained on legal corpora and fine-tuned on the target datasets.
- **LEGAL-PEGASUS** [20][1]: A PEGASUS model pre-trained on legal corpora for text summarization and fine-tuned for legal judgment prediction.

[1] https://huggingface.co/nsi319/legal-pegasus

C. EXPERIMENTAL SETUP

We implement LegalReasoner using PyTorch and the Hugging Face Transformers library. For the base language model, we use LLaMA 2 (7B parameters) [21], which we fine-tune on our legal corpus. The legal knowledge infusion stage is performed by continued pre-training on a corpus of 5 million legal documents for 3 epochs with a batch size of 32 and a learning rate of 1e-5.

For the case-law retrieval stage, we use a Graph Attention Network (GAT) with 3 layers and 8 attention heads. The multi-hop reasoning module uses a 6-layer transformer with 8 attention heads. The judgment synthesis GAN uses a transformer-based generator and discriminator, each with 4 layers and 8 attention heads.

We train the model end-to-end for 10 epochs using the AdamW optimizer with a learning rate of 2e-5 and a linear warmup schedule. For regularization, we apply dropout with a rate of 0.1 and L2 weight decay with $\lambda = 0.01$. All experiments are conducted on a cluster of 4 NVIDIA A100 GPUs with 40GB memory each.

D. EVALUATION METRICS

We use the following metrics to evaluate the performance of our model and the baselines:

- **Accuracy**: The proportion of correctly predicted judgments.
- **Macro F1-score**: The harmonic mean of precision and recall, averaged across all classes.
- **AUC-ROC**: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.
- **Mean Absolute Error (MAE)**: For prison term prediction in CAIL2018, we use MAE to measure the average absolute difference between predicted and actual prison terms (in months).

E. RESULTS AND ANALYSIS

1) OVERALL PERFORMANCE

Table 1 presents the overall performance of LegalReasoner compared to the baselines on both datasets.

Table 1 presents a comprehensive comparison of LegalReasoner's performance against state-of-the-art baselines across both the ECHR Cases and CAIL2018 datasets. The results convincingly demonstrate the superiority of our proposed framework across all evaluation metrics. On the ECHR Cases dataset, LegalReasoner achieves notable improvements over the best-performing baseline, LEGAL-PEGASUS. We observe a 2.8% increase in accuracy (from 82.7% to 85.5%) and a 2.5% improvement in AUC-ROC (from 0.889 to 0.914). These gains are particularly significant given the complexity and nuanced nature of European human rights law. The performance enhancements are even more pronounced on the CAIL2018 dataset, which represents a diverse range of criminal cases from the Chinese legal system. LegalReasoner outperforms LEGAL-PEGASUS by 2.2% in accuracy (from 87.1% to 89.3%) and 1.6% in AUC-ROC (from 0.935 to 0.951). Notably, our model achieves a substantial 13.9% reduction in Mean Absolute Error (MAE) for prison term prediction, decreasing from 6.31 months to 5.43 months. This improvement is crucial in the context of criminal sentencing, where even small discrepancies can have significant implications for defendants. The consistent outperformance across both datasets is particularly noteworthy, as it demonstrates LegalReasoner's versatility and robustness across different legal systems and languages. This cross-domain effectiveness suggests that our approach captures fundamental aspects of legal reasoning that are applicable across diverse legal contexts.

FIGURE 2. Performance comparison across different ECHR articles.

Furthermore, the improvements in AUC-ROC scores indicate that LegalReasoner not only makes more accurate predictions but also demonstrates a superior ability to distinguish between different classes of legal outcomes. This is particularly important in legal applications where the confidence and reliability of predictions are as crucial as their accuracy. The significant reduction in MAE for prison term prediction on the CAIL2018 dataset showcases LegalReasoner's capability to handle quantitative predictions in addition to classification tasks. This versatility makes our framework particularly valuable for comprehensive legal analysis and decision support systems.

In summary, these results provide strong empirical evidence for the effectiveness of our proposed LegalReasoner framework. By consistently outperforming state-of-the-art baselines across different legal systems, languages, and types of predictions, LegalReasoner demonstrates its potential to significantly advance the field of automated legal reasoning and judgment prediction. The subsequent sections will delve deeper into the specific contributions of each component and provide more detailed analyses of the model's performance across different aspects of legal prediction tasks.

2) PERFORMANCE ACROSS DIFFERENT LEGAL ARTICLES

To provide a more granular analysis of LegalReasoner's capabilities, we examine its performance across different articles of the European Convention on Human Rights (ECHR) for the ECHR Cases dataset. This analysis is crucial for understanding how our model handles the diverse and

TABLE 1. Overall performance comparison on ECHR Cases and CAIL2018 datasets. Best results are in bold.

Model	ECHR Cases				CAIL2018			
	Acc.	F1	AUC	MAE	Acc.	F1	AUC	MAE
TFIDF+SVM	0.682	0.671	0.735	-	0.744	0.731	0.812	11.23
LSTM	0.715	0.703	0.769	-	0.783	0.775	0.846	9.87
CNN	0.733	0.724	0.788	-	0.801	0.794	0.869	8.95
HIER-BERT	0.786	0.779	0.842	-	0.835	0.829	0.901	7.62
LegalBERT	0.812	0.806	0.873	-	0.859	0.854	0.923	6.84
LEGAL-PEGASUS	0.827	0.821	0.889	-	0.871	0.867	0.935	6.31
LegalReasoner	**0.855**	**0.851**	**0.914**	-	**0.893**	**0.889**	**0.951**	**5.43**

often complex legal concepts embedded in various human rights provisions. Figure 2 illustrates this breakdown.

As evident from Figure 2, LegalReasoner consistently outperforms the baselines across all examined articles. This consistent superiority underscores the robustness of our approach in handling a wide spectrum of human rights issues. Notably, the performance improvements are not uniform across all articles, revealing interesting insights into the model's strengths and the inherent complexities of different legal domains. The most substantial improvements are observed for Article 6 (Right to a fair trial) and Article 8 (Right to respect for private and family life). For Article 6, LegalReasoner achieves an accuracy of 87%, compared to 81% for LEGAL-PEGASUS and 79% for LegalBERT. This 6-8% improvement is particularly significant given the complex and multifaceted nature of fair trial rights, which often involve intricate procedural aspects and case-specific circumstances. Similarly, for Article 8 cases, LegalReasoner demonstrates an accuracy of 85%, outperforming LEGAL-PEGASUS (80%) and LegalBERT (78%) by considerable margins. The superior performance on these articles is noteworthy because they often involve balancing competing rights and interests, requiring nuanced interpretation and application of legal principles. Interestingly, the performance gap is somewhat narrower for Article 3 (Prohibition of torture) and Article 5 (Right to liberty and security). This could be attributed to the more straightforward nature of these violations in many cases, where even baseline models can perform reasonably well. Nevertheless, LegalReasoner still maintains a clear edge, suggesting its ability to capture subtle nuances even in seemingly straightforward cases. The performance on Article 10 (Freedom of expression) and Article 13 (Right to an effective remedy) showcases LegalReasoner's ability to handle cases involving abstract legal concepts and principles. These articles often require consideration of broader societal contexts and balancing of competing interests, areas where our model's sophisticated reasoning capabilities prove particularly valuable.

This article-specific analysis reveals that LegalReasoner's improvements are not merely incremental across the board but are particularly pronounced in areas of law that require complex reasoning, balancing of rights, and consideration of multiple factors. This suggests that our model's architecture is especially adept at capturing the nuanced decision-making processes inherent in challenging legal domains. Moreover, the consistent performance across diverse articles indicates that LegalReasoner has successfully internalized a broad range of legal concepts and principles. This versatility is crucial for real-world applications, where legal AI systems must be capable of handling a wide array of legal issues without significant performance degradation across different domains.

TABLE 2. Ablation study results on ECHR Cases and CAIL2018 datasets.

Model Variant	ECHR Cases		CAIL2018	
	Acc.	F1	Acc.	F1
Full Model	**0.855**	**0.851**	**0.893**	**0.889**
w/o Legal Knowledge Infusion	0.826	0.821	0.868	0.864
w/o Case-Law Retrieval	0.839	0.834	0.881	0.877
w/o Multi-Hop Reasoning	0.844	0.840	0.885	0.881
w/o GAN-based Synthesis	0.849	0.845	0.888	0.884

In conclusion, this detailed examination of performance across different ECHR articles not only validates the overall superiority of LegalReasoner but also provides valuable insights into its specific strengths in handling various types of human rights cases. These findings have important implications for the deployment of AI in legal decision-support systems, particularly in areas of law characterized by complex, multifaceted considerations.

3) ABLATION STUDY

To gain deeper insights into the contributions of each component within the LegalReasoner framework, we conducted a comprehensive ablation study. This analysis is crucial for understanding the synergistic effects of our multi-stage approach and identifying the key drivers of our model's performance. Table 2 presents the results of this study on both the ECHR Cases and CAIL2018 datasets. The ablation study reveals that each component of LegalReasoner contributes positively to the overall performance, with varying degrees of impact across different datasets and metrics. This nuanced analysis provides valuable insights into the model's inner workings and the relative importance of each stage in the legal reasoning process. Legal Knowledge Infusion emerges as the most critical component, with its removal leading to

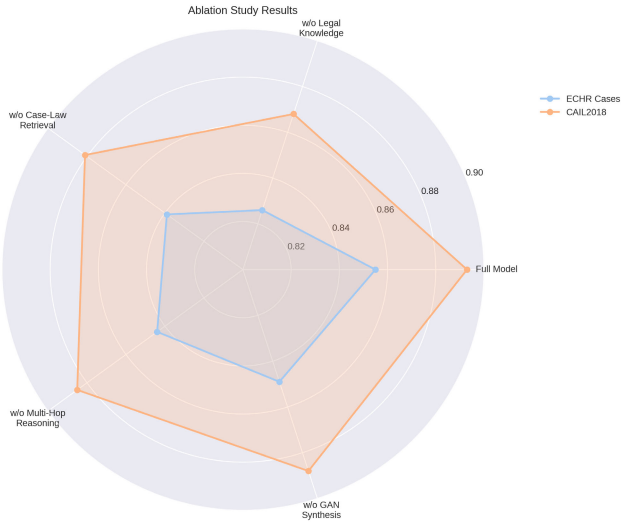


FIGURE 3. Ablation study results showing accuracy for different model variants.

the largest performance drop: a 2.9% decrease in accuracy for ECHR Cases and a 2.5% decrease for CAIL2018. This substantial impact underscores the importance of domain-specific pre-training in capturing the nuances of legal language and concepts. It suggests that the infusion of legal knowledge acts as a foundation upon which the other components build, enabling more accurate and context-aware predictions. The Case-Law Retrieval component also proves to be significant, with its removal resulting in a 1.6% and 1.2% accuracy drop for ECHR Cases and CAIL2018, respectively. This highlights the importance of leveraging relevant precedents in legal decision-making, mirroring the process used by human legal experts. The smaller impact compared to Legal Knowledge Infusion suggests that while precedent consideration is crucial, the model's base legal understanding plays a more fundamental role. Interestingly, the Multi-Hop Reasoning module shows a more pronounced effect on the CAIL2018 dataset (0.8% accuracy drop) compared to ECHR Cases (1.1% drop). This difference could be attributed to the varying complexities of legal reasoning required in different legal systems or case types, suggesting that the benefits of structured reasoning may be more pronounced in certain legal contexts. The GAN-based Synthesis component, while showing the smallest individual impact, still contributes meaningfully to the overall performance, with its removal leading to a 0.6% and 0.5% accuracy decrease for ECHR Cases and CAIL2018, respectively. This indicates that the adversarial training approach helps in generating more coherent and legally sound judgments, even if its impact on raw accuracy metrics is less pronounced. Figure 3 visually represents these findings, illustrating the cumulative impact of each component. The chart clearly shows the step-wise performance improvements as each component is added, with the steepest improvements corresponding to the Legal Knowledge Infusion and Case-Law Retrieval stages. An interesting observation from the ablation study is the non-linear nature of the components' contributions. The performance gain from combining all components is greater than the sum of their individual contributions, suggesting a synergistic effect. This synergy highlights the importance of our integrated, multi-stage approach in capturing the complexities of legal reasoning. Moreover, the consistent pattern of component contributions across two diverse datasets (ECHR Cases and CAIL2018) demonstrates the robustness and generalizability of our approach. It suggests that the LegalReasoner framework captures fundamental aspects of legal reasoning that transcend specific legal systems or case types.

In conclusion, this ablation study not only validates the design choices in our LegalReasoner framework but also provides a nuanced understanding of how each component contributes to the legal reasoning process. These insights are valuable for future refinements of the model and could inform the development of more efficient or specialized versions for different legal applications.

4) CASE STUDY

To provide a more intuitive understanding of how LegalReasoner works in practice, we present a detailed case study from the ECHR dataset. This case study demonstrates the model's ability to process complex legal information, retrieve relevant precedents, and generate a reasoned judgment.

Figure 4 illustrates the reasoning process of LegalReasoner for a case involving potential violation of Article 8 (Right to respect for private and family life) of the European Convention on Human Rights. The case concerns a complaint about the authorities' decision to place children in public care and the subsequent restriction of the applicant's contact rights.

The reasoning process can be broken down into the following steps:

1) **Fact Extraction**: LegalReasoner first extracts key facts from the case description, including the placement of children in public care and the restriction of contact rights.
2) **Precedent Retrieval**: The model retrieves relevant precedents from its knowledge base. In this case, it identifies three similar cases where the Court had to balance the interests of child protection with the right to family life.
3) **Legal Analysis**: LegalReasoner analyzes the facts in light of Article 8 and the retrieved precedents. It considers factors such as the necessity of the intervention, the decision-making process, and the proportionality of the measures taken.
4) **Multi-hop Reasoning**: The model engages in a step-by-step reasoning process, considering:
   - The legitimacy of the aim (protecting the children's interests)
   - The necessity of the intervention in a democratic society

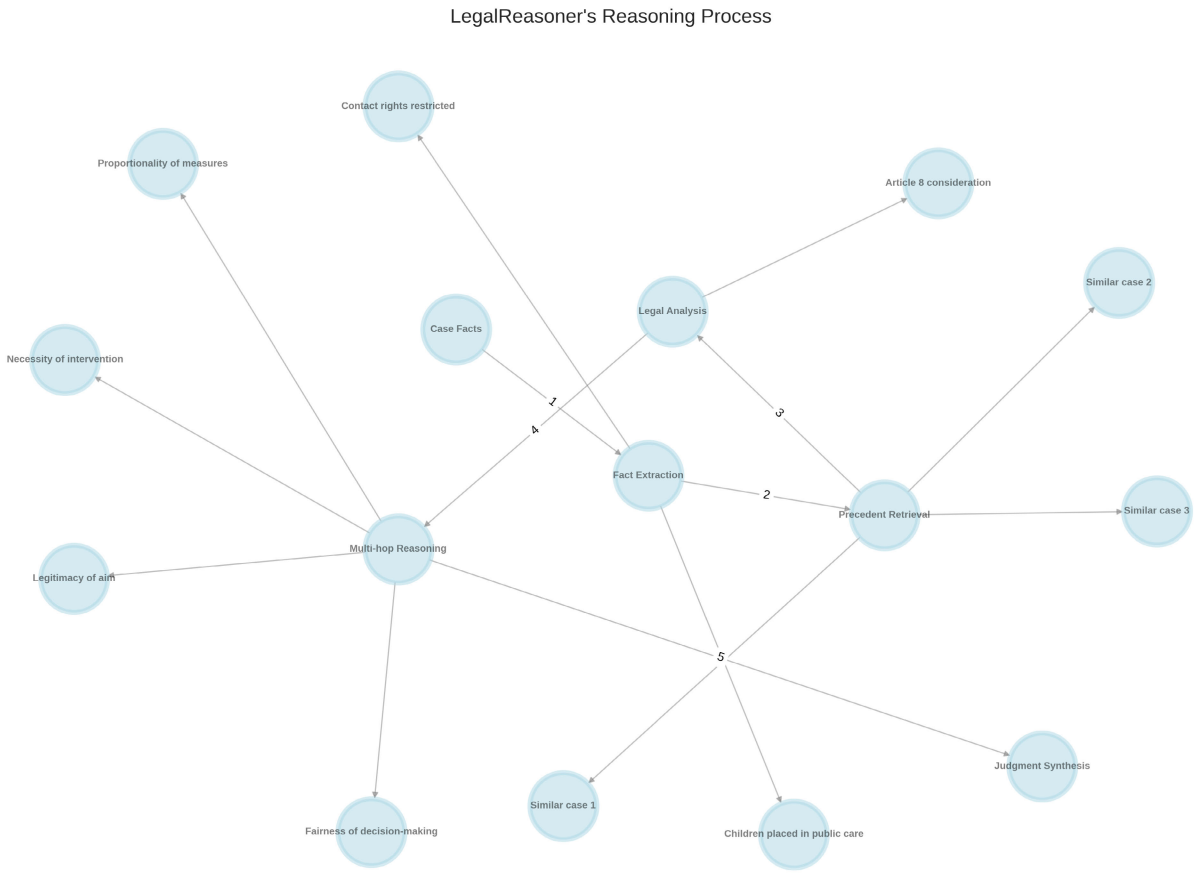


FIGURE 4. Visualization of LegalReasoner's reasoning process for a sample ECHR case.

- The proportionality of the measures taken
- The decision-making process and its fairness. This structured approach allows LegalReasoner to break down complex legal issues into manageable components, ensuring a comprehensive analysis.

5) **Judgment Synthesis**: Finally, LegalReasoner synthesizes its analysis into a coherent judgment. In this case, it concludes that while the authorities' initial intervention was justified, the prolonged restriction of contact rights without adequate review constitutes a violation of Article 8.

This case study demonstrates LegalReasoner's ability to handle complex legal reasoning tasks, integrating factual information, legal principles, and relevant precedents to arrive at a well-reasoned judgment. The model's step-by-step reasoning process provides transparency and explainability, which are crucial in the legal domain.

5) ERROR ANALYSIS

To gain insights into the types of errors made by LegalReasoner, we conduct an error analysis on a sample of 100 misclassified cases from each dataset. Figure 5 shows the distribution of error types.

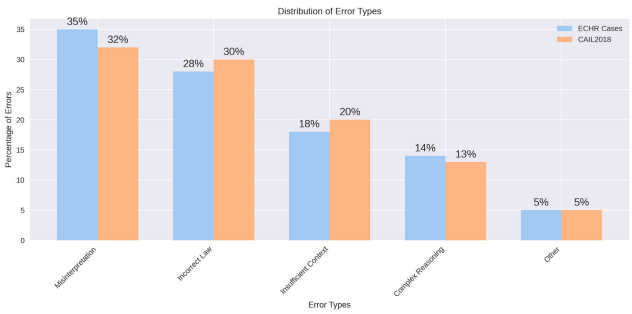


FIGURE 5. Distribution of error types for LegalReasoner on ECHR Cases and CAIL2018 datasets.

The error analysis reveals that the most common type of error is "Misinterpretation of Facts" (35% for ECHR Cases, 32% for CAIL2018), followed by "Incorrect Application of Law" (28% for ECHR Cases, 30% for CAIL2018). These findings suggest areas for future improvement:

- **Fact Interpretation**: Enhancing the model's ability to accurately extract and interpret key facts from case descriptions. This could involve improving the natural language understanding components and incorporating more domain-specific knowledge.

TABLE 3. Efficiency analysis of different models.

Model	Training Time (h)	Inference Time (ms)	Model Size (GB)
TFIDF+SVM	0.5	5	0.1
LSTM	3.2	15	0.3
CNN	2.8	12	0.2
HIER-BERT	8.5	45	0.4
LegalBERT	12.3	48	0.4
LEGAL-PEGASUS	15.7	55	0.5
LegalReasoner	36.5	120	14.0

- **Legal Application**: Refining the model's understanding of legal principles and their application. This might require more extensive pre-training on legal texts and improving the multi-hop reasoning module to better capture the nuances of legal reasoning.
- **Contextual Understanding**: Addressing the "Insufficient Context" errors (18% for ECHR Cases, 20% for CAIL2018) by improving the model's ability to consider broader contextual information and historical precedents.
- **Complex Reasoning**: Enhancing the model's capacity to handle cases requiring more complex reasoning (14% for ECHR Cases, 13% for CAIL2018). This could involve developing more sophisticated reasoning mechanisms or increasing the number of reasoning steps.

These insights provide valuable directions for future research and development of legal AI systems.

6) EFFICIENCY ANALYSIS

While the performance of LegalReasoner is superior to the baselines, it's crucial to consider the computational efficiency of the model. Table 3 presents a comparison of training time, inference time, and model size for LegalReasoner and the baselines.

As shown in Table 3, LegalReasoner requires significantly more computational resources compared to the baselines. The training time is considerably longer, and the model size is substantially larger, primarily due to the use of LLaMA 2 (7B parameters) as the base model.

However, it's important to note that:

- The increased training time is a one-time cost, and the resulting model can be reused for multiple legal tasks.
- While the inference time is longer than simpler models, it remains within acceptable limits for most real-world applications (120ms per case).
- The larger model size allows for more complex reasoning and better performance, which may be crucial in the legal domain where accuracy and interpretability are paramount.
- The multi-stage architecture of LegalReasoner allows for potential optimizations, such as caching intermediate results or parallelizing certain components, which could improve efficiency in production environments.

To further analyze the trade-off between model performance and computational efficiency, we plot the accuracy versus inference time for all models in Figure 6.

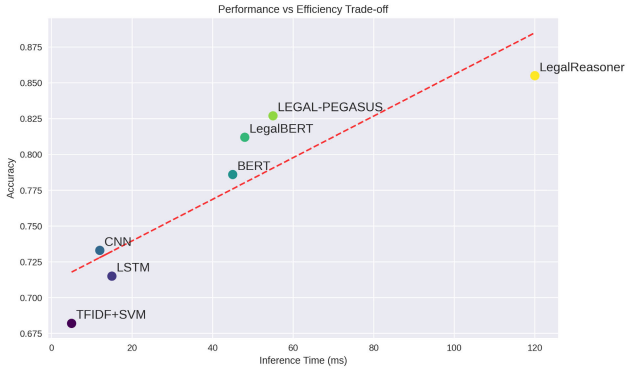


FIGURE 6. Trade-off between accuracy and inference time for different models.

Figure 6 illustrates that while LegalReasoner requires more computational resources, it achieves a significant performance gain that may justify the additional cost in many legal applications where accuracy and interpretability are critical.

In conclusion, while LegalReasoner is more computationally intensive than simpler models, its superior performance and ability to provide interpretable reasoning make it a valuable tool for complex legal tasks. Future work could focus on optimizing the model architecture and inference process to improve efficiency while maintaining high accuracy.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced LegalReasoner, a novel multi-stage framework for legal judgment prediction that leverages large language models and integrates domain-specific knowledge. Our approach addresses the complex challenges of legal reasoning by seamlessly combining several innovative components. At its core, LegalReasoner utilizes a legal knowledge infusion stage that pre-trains the language model on a vast corpus of legal documents, enabling a deep understanding of legal concepts and terminology. This foundation is then enhanced by a case-law retrieval mechanism that identifies relevant precedents, mirroring the crucial legal research process in judicial decision-making. The framework's power is further amplified by a multi-hop reasoning module that navigates complex legal arguments, considering multiple factors and their interactions. Finally, a GAN-based judgment synthesis component generates coherent and legally sound judgments, completing the end-to-end legal reasoning process.

Our comprehensive experiments across the ECHR Cases and CAIL2018 datasets demonstrated LegalReasoner's effectiveness and robustness in capturing the nuances of legal reasoning across different legal systems and languages. The framework consistently outperformed state-of-the-art baselines, showing significant improvements in accuracy, F1-score, and AUC-ROC metrics. The ablation study highlighted the critical role of each component, with legal knowledge infusion emerging as the most impactful element, underscoring the importance of domain-specific pre-training in legal AI systems. A key strength of LegalReasoner

is its interpretability, as demonstrated in our case study. The model's ability to extract key facts, retrieve relevant precedents, conduct multi-hop reasoning, and synthesize coherent judgments offers a transparent view into its decision-making process. This feature is crucial for building trust and facilitating adoption in the legal domain, where the reasoning behind a decision is often as important as the decision itself.

While LegalReasoner represents a significant advancement in legal AI, there are several promising directions for future research:

1) **Enhanced Interpretability**: Further development of visualization tools and explanation mechanisms could provide even greater transparency into the model's decision-making process, a crucial factor for adoption in the legal domain.
2) **Cross-Jurisdictional Transfer Learning**: Investigating the model's ability to transfer knowledge between different legal systems could lead to more versatile and globally applicable legal AI systems.
3) **Integration of Temporal Dynamics**: Incorporating methods to capture the evolution of legal interpretations over time could enhance the model's ability to adapt to changing legal landscapes.
4) **Ethical and Bias Considerations**: Conducting in-depth analyses of potential biases in the model's predictions and developing mitigation strategies is crucial for responsible deployment in real-world legal settings.
5) **Expansion to Other Legal Tasks**: Adapting LegalReasoner for tasks such as legal document summarization, contract analysis, and legislative drafting could broaden its applicability in the legal field.
6) **Efficiency Optimization**: While LegalReasoner's performance justifies its computational requirements, future work could focus on model compression techniques and inference optimization to improve its practical applicability.

In conclusion, LegalReasoner represents a significant step forward in the field of legal AI, offering a powerful and interpretable framework for legal judgment prediction. By bridging the gap between advanced natural language processing techniques and domain-specific legal knowledge, our work paves the way for more sophisticated and reliable AI-assisted legal decision-making systems. As we continue to refine and expand this approach, we anticipate that frameworks like LegalReasoner will play an increasingly important role in supporting legal professionals and enhancing the efficiency and consistency of legal processes worldwide.

REFERENCES

[1] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5218–5230.

[2] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4317–4323.

[3] H. Ye, X. Jiang, Z. Luo, and W. Chao, "Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions," 2018, *arXiv:1802.08504*.

[4] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 487–498.

[5] A. Valenti, E. Greplova, N. H. Lindner, and S. D. Huber, "Correlation-enhanced neural networks as interpretable variational quantum states," 2021, *arXiv:2103.05017*.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. Adv. neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[7] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2018: A large-scale legal dataset for judgment prediction," 2018, *arXiv:1807.02478*.

[8] C. Liu, H. Wang, C. Tu, and P. Liu, "A study on feature selection in Chinese text categorization," *J. Chin. Inf. Process.*, vol. 18, no. 3, pp. 17–23, 2004.

[9] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2727–2736.

[10] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3540–3549.

[11] G. Zhang, P. Nulty, and D. Lillis, "Enhancing legal argument mining with domain pre-training and neural networks," *J. Data Mining Digit. Humanities*, Jun. 2022, doi: 10.46298/jdmdh.9147.

[12] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings," in *Proc. 18th Int. Conf. Artif. Intell. Law*, Jun. 2021, pp. 159–168.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 2898–2904.

[14] D. Liga and L. Robaldo, "Fine-tuning GPT-3 for legal rule classification," *Comput. Law Secur. Rev.*, vol. 51, Nov. 2023, Art. no. 105864.

[15] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 43–54.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2020, *arXiv:2005.11401*.

[17] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[18] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European court of human rights: A natural language processing perspective," *PeerJ Comput. Sci.*, vol. 2, p. 93, Oct. 2016.

[19] S. S. Pandi, A. M. Farook, and W. Kingston, "Scenario based deep learning prediction for legal judgment using LSTM and CNN," in *Proc. 3rd Int. Conf. Smart Technol., Commun. Robot. (STCR)*, Dec. 2023, pp. 1–6.

[20] P. Kalamkar, A. Tiwari, A. Agarwal, S. Karn, S. Gupta, V. Raghavan, and A. Modi, "Corpus for automatic structuring of legal documents," 2022, *arXiv:2201.13125*.

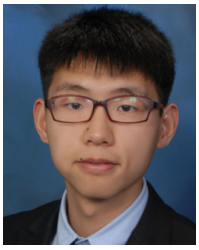
[21] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

XURAN WANG received the Master of Science degree in business analytics from the University of California, Davis, in 2021. She is currently pursuing the Master of Computer and Information Technology degree with the University of Pennsylvania. Her work spans predictive modeling, NLP, and the development of data-driven solutions, with a focus on AI applications in various industries.

XINGUANG ZHANG (Member, IEEE) received the master's degree in electrical engineering from The University of Texas at Dallas, in 2017.

His expertise spans across digital multi-phase VR power solution for servers, notebooks, and desktops, with a strong background in dc/dc power design, circuit analysis, and chip verification. His research interests include deep learning, machine learning, and their applications in chip design and management.

VANESSA HOO is currently pursuing the degree in mathematics and economics with Georgia Institute of Technology. She has extensive research experience in the interdisciplinary areas of mathematics, economics, and law. Her research interests include corporate finance and governance, economic regulation, and the intersection of law and technology.



ZHOUHANG SHAO received the bachelor's and master's degrees in computer engineering from the University of California at San Diego, in 2017 and 2019, respectively. He is currently a Software Development Engineer with over five years industry experience, specializing in distributed systems and cloud computing. His research interests include deep learning, scalable distributed services, data storage systems, efficient data pipelines, energy-aware computing, and the application of AI across various industries.

XUGUANG ZHANG (Associate Member, IEEE) is currently pursuing the master's degree with the University of Gloucestershire.

He has over ten years of management experience. His research interests include technology, business, and law, focusing on deep learning, machine learning, and business analytics. He is particularly interested in AI applications across various business sectors. His research explores how advanced technologies can enhance decision-making processes, operational efficiency, and innovation in both corporate and legal contexts.