

Article

Not peer-reviewed version

LawLLM-DS: A Two-Stage Parameter-Efficient Fine-Tuning Framework for Legal Judgment Prediction with Symmetry-Aware Label Graphs

[Pengcheng Zhao](#)[†], [Chengcheng Han](#), [Kun Han](#)^{*†}

Posted Date: 17 December 2025

doi: 10.20944/preprints202512.1406.v1

Keywords: legal judgment prediction; large language models; LoRA; parameter-efficient fine-tuning; two-stage training; symmetry-aware label graphs; graph-based representations; domain adaptation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

LawLLM-DS: A Two-Stage Parameter-Efficient Fine-Tuning Framework for Legal Judgment Prediction with Symmetry-Aware Label Graphs

Pengcheng Zhao ^{1,†}, Chengcheng Han ² and Kun Han ^{1,*,†}

¹ Faculty of Law, Macau University of Science and Technology, Macau 999078, China

² Faculty of Economics, University of New South Wales, Sydney, NSW 2033, Australia

* Correspondence: 3230004066@student.must.edu.mo

† These authors contributed equally to this work.

Abstract

Legal judgment prediction (LJP) increasingly relies on large language models whose full fine-tuning is memory-intensive and susceptible to catastrophic forgetting. We present LawLLM-DS, a two-stage Low-Rank Adaptation (LoRA) framework that first performs legal knowledge pre-tuning with an aggressive learning rate and subsequently refines judgment relations with conservative updates, using dedicated LoRA adapters, 4-bit quantization, and targeted modification of seven Transformer projection matrices to keep only 0.21% of parameters trainable. From a structural perspective, the twenty annotated legal elements form a symmetric label co-occurrence graph that exhibits both cluster-level regularities and asymmetric sparsity patterns, and LawLLM-DS implicitly captures these graph-informed dependencies while remaining compatible with downstream GNN-based representations. Experiments on 5,096 manually annotated divorce cases show that LawLLM-DS lifts macro F1 to 0.8893 and achieves an accuracy of 0.8786, outperforming single-stage LoRA and BERT baselines under the same data regime. Ablation studies further verify the contributions of stage-wise learning rates, adapter placement, and low-rank settings. These findings demonstrate that curriculum-style, parameter-efficient adaptation provides a practical path toward lightweight yet structure-aware LJP systems for judicial decision support.

Keywords: legal judgment prediction; large language models; LoRA; parameter-efficient fine-tuning; two-stage training; symmetry-aware label graphs; graph-based representations; domain adaptation

1. Introduction

Legal judgment prediction has emerged as a critical application domain at the intersection of artificial intelligence and legal informatics, with profound implications for judicial decision support systems, legal consultation services, and case outcome assessment [1–3]. The task involves predicting judicial outcomes based on textual descriptions of case facts, requiring sophisticated understanding of legal concepts, factual elements, and their interrelationships within the judicial reasoning framework. In multi-label civil judgment prediction—such as marital property allocation, child custody arrangements, and divorce grounds—these legal elements naturally form a structured system of dependencies. When viewed through a graph-theoretic lens, each case and its annotated labels constitute a case–label bipartite graph, whose induced label co-occurrence graph exhibits cluster-level regularities as well as asymmetric sparsity patterns across different legal factors. Such structures motivate learning approaches that are able to respect both the symmetric and asymmetric patterns inherent in legal judgment graphs, and that remain compatible with downstream graph-based reasoning or GNN architectures.

Beyond strictly predictive tasks, adjacent strands of legal NLP research investigate long-document summarization pipelines for case law [4], Arabic judicial decision support platforms that integrate

retrieval with classification [5], and large-scale crime prediction surveys that contextualize machine learning deployments within policing workflows [6,7]. Hybrid deep neural frameworks tailored for multi-task legal analytics [8] further illustrate the diversity of architectures required to reason over procedural posture, statutory factors, and evidentiary descriptions. However, these models often overlook the structural regularities arising from the graph organization of legal elements and their co-occurrence patterns, which contain both symmetric correlations within thematic clusters and asymmetric dependencies associated with rare or sparsely connected labels.

Recent advances in deep learning have significantly improved legal judgment prediction through sophisticated text representation and classification techniques. Traditional machine learning approaches including support vector machines, random forests, and gradient boosting methods [9,10] rely on handcrafted features and thus struggle to capture the deeper structural dependencies across labels that naturally emerge from the underlying legal judgment graph. Deep neural architectures including convolutional neural networks [11], recurrent neural networks with LSTM and GRU cells [12], and attention-based models [13] are able to discover hierarchical textual patterns automatically, yet they still lack explicit mechanisms to incorporate the symmetric or asymmetric label interactions encoded in the label graph. Furthermore, these task-specific architectures must be trained from scratch and often require substantial labeled data and computational resources, limiting their applicability to domains such as family law where curated datasets are expensive to obtain.

Pre-trained language models have revolutionized natural language processing by enabling effective transfer learning from large-scale unsupervised corpora to downstream tasks. Models such as BERT [14] and domain-adapted variants [15] capture rich linguistic and semantic knowledge during pre-training and have demonstrated strong performance in legal text analysis. However, standard full-parameter fine-tuning faces several critical challenges. First, updating all model parameters is computationally expensive in terms of memory, storage, and training time. Second, full fine-tuning increases the risk of catastrophic forgetting, potentially disrupting previously learned linguistic or structural knowledge [16]. Third, uniform learning rate schedules may fail to reconcile broad domain adaptation objectives with the fine-grained optimization required to capture asymmetric judgment patterns across labels—patterns which, from a graph-structural perspective, correspond to highly imbalanced or weakly connected regions of the label co-occurrence graph.

Large language models with billions of parameters [17,18] encode extensive world knowledge, including legal terminology and reasoning heuristics, but adapting them reliably to specialized legal tasks remains computationally challenging. Recent applications have explored LLMs for legal question answering, contract analysis, and judgment prediction [19–21], while broader scholarship emphasizes responsible governance in emerging socio-technical domains [22]. The rapid diffusion of LLMs into safety-critical environments such as neurosurgical simulation [23], autonomous control [24], and digital forensics [25] further underscores the need for robust, well-structured fine-tuning strategies. Although prompt engineering offers a lightweight alternative, it is often insufficient for complex multi-label classification tasks that require nuanced modeling of the structured, graph-like dependencies among legal elements [26]. These limitations create an opportunity for parameter-efficient fine-tuning methods that can capture both textual and structural patterns in legal judgment graphs while remaining computationally affordable.

Parameter-efficient fine-tuning methods have emerged as a promising solution to adapt large pre-trained models while maintaining computational efficiency. Low-Rank Adaptation (LoRA) represents a breakthrough in this direction by constraining weight updates to low-rank subspaces, reducing trainable parameters to a small fraction of the original model while achieving performance comparable to full fine-tuning [27]. The fundamental insight underlying LoRA is that pre-trained language models exhibit intrinsically low-dimensional structures, a property that resonates with the symmetry patterns commonly observed in graph-structured data. When legal judgments are represented through a case-label bipartite graph, or through its induced symmetric label co-occurrence matrix, many legal elements display clustered, approximately symmetric correlations, whereas rare labels form highly asymmetric,

sparsely connected regions. LoRA's low-rank parameterization provides a natural mechanism for adapting the model along these dominant symmetric directions while avoiding overfitting to noisy or weakly connected asymmetric substructures of the legal judgment graph.

By decomposing weight updates into low-rank matrices, LoRA enables efficient domain adaptation while preserving the linguistic knowledge encoded during pre-training. Despite its effectiveness, existing applications of LoRA primarily employ single-stage fine-tuning with uniform learning rates, potentially overlooking the structural asymmetry inherent in legal label graphs and missing opportunities for progressive knowledge acquisition through curriculum-inspired training strategies. In particular, symmetric clusters of frequently co-occurring labels may be learned rapidly, whereas asymmetric, low-frequency legal elements require more conservative refinement to avoid gradient instability and catastrophic forgetting.

To address these limitations and harness the full potential of large language models for legal judgment prediction while maintaining parameter efficiency, this paper proposes a two-stage sequential fine-tuning framework based on LoRA that decomposes the learning process into progressive knowledge acquisition stages with differentiated learning rates. The key insight is that legal judgment prediction requires both broad legal domain knowledge—corresponding to learning the approximately symmetric structure of the label co-occurrence graph—and fine-grained relationship modeling between specific factual elements and judicial outcomes, which often reside in asymmetric or sparsely connected regions of the graph. By separating these learning objectives into distinct sequential stages with appropriately calibrated learning rates, the proposed approach enables the model to first establish comprehensive domain understanding through rapid exploration of symmetric patterns, followed by refined optimization of decision boundaries through conservative updates that stabilize learning on asymmetric label structures.

The main contributions of this work are:

- This paper proposes a two-stage sequential fine-tuning framework based on Low-Rank Adaptation (LoRA) that decomposes legal judgment prediction into progressive knowledge acquisition stages. Stage 1 performs legal knowledge pre-tuning with a higher learning rate ($\eta_1 = 3 \times 10^{-4}$), enabling the model to rapidly capture dominant, approximately symmetric structures in the label co-occurrence graph, while Stage 2 conducts judgment relation fine-tuning with a lower learning rate ($\eta_2 = 1 \times 10^{-4}$) that refines asymmetric and low-frequency label relationships, achieving efficient transfer learning with only 0.21% trainable parameters.
- This work presents a comprehensive analysis of the theoretical foundations and practical implementation of LoRA-based parameter-efficient fine-tuning for legal text classification. By aligning low-rank parameterization with the structural regularities of the legal judgment graph, the framework provides a symmetry-aware adaptation strategy, along with detailed formulations of optimization objectives and gradient computation procedures for both training stages.
- Extensive experiments are conducted on a Chinese legal case dataset comprising divorce-related judgments with 20 multi-label categories. The results demonstrate that the proposed two-stage approach achieves superior performance compared to single-stage fine-tuning baselines across multiple evaluation metrics including macro F1-score, Hamming Loss, and accuracy, validating that separating symmetric and asymmetric learning behaviors through curriculum-inspired sequential optimization is effective for complex legal domain adaptation tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on legal judgment prediction, graph-based representations, and parameter-efficient fine-tuning techniques for large language models. Section 3 presents the detailed methodology including problem formulation, the graph-theoretic interpretation of legal label structure, LoRA fundamentals, the proposed two-stage sequential fine-tuning strategy, and implementation details. Section 4 describes the experimental setup, dataset characteristics, evaluation metrics, and presents comprehensive results including performance comparisons, ablation studies, and hyperparameter sensitivity analysis. Section 5 concludes the

paper with a summary of key findings and discusses future research directions for structure-aware, parameter-efficient adaptation of large language models in legal and other graph-structured domains.

1.1. Legal Judgment Prediction

Legal judgment prediction has attracted significant research attention in legal artificial intelligence, evolving from traditional machine learning approaches to sophisticated neural architectures. Early work employed feature engineering combined with classifiers including support vector machines [9,28], random forests [10], and logistic regression, achieving reasonable performance but requiring substantial domain expertise for feature design. These methods typically extract statistical features, n-gram representations, and manually crafted legal concept indicators from case documents, but they do not explicitly model the structural dependencies among legal elements that naturally arise in case-label bipartite graphs or label co-occurrence graphs.

Deep learning approaches have demonstrated superior performance by automatically learning hierarchical representations from legal texts. Convolutional neural networks [11] extract n-gram patterns at multiple scales through hierarchical convolution and pooling operations. Recurrent neural networks including LSTM and GRU variants [12,29] capture sequential dependencies in legal narratives through recurrent connections. Attention-based architectures [13,27,30] enable models to focus on salient textual segments relevant to specific legal elements. More recently, graph neural networks (GNNs) [2,31] have been introduced to model relationships among legal articles, precedents, and factual elements by treating them as nodes in a heterogeneous graph whose edges encode citation, semantic, or co-occurrence relations. These approaches leverage advances in graph connectivity theory [32–34] and network diagnosability [35–38], enabling reasoning over both symmetric structural clusters and asymmetric sparse connections—a perspective highly relevant to multi-label divorce judgments where legal elements exhibit clustered co-occurrence patterns alongside rare, weakly connected labels.

Legal-specific hybrid architectures including MPBFN [39], LADAN [40], and Mulan [41] integrate multiple attention heads or feedback pathways to distinguish confusing statutes under multi-label supervision, while reinforcement learning and rationale extraction frameworks [42,43] emphasize interpretable identification of decisive criminal elements. Knowledge-graph enhanced encoders [44] further demonstrate the benefit of jointly modeling semantic and topological cues, underscoring the importance of representing legal elements not only as independent labels but as nodes embedded in a graph structure that exhibits both symmetry and asymmetry.

Pre-trained language models have achieved state-of-the-art performance across legal NLP tasks. BERT [14] and its variants provide contextualized representations through bidirectional Transformer encoders trained on masked language modeling objectives. Domain-specific variants and specialized architectures have been developed through continued pre-training on legal corpora, achieving enhanced performance on legal text understanding tasks [15,45]. Recent work explores multi-task learning [46–48], contrastive learning [49], and graph-enhanced methods [31] to further improve prediction accuracy by exploiting structural regularities in legal judgment graphs.

However, most existing approaches require full fine-tuning of model parameters, leading to substantial computational costs when adapting large pre-trained models. Moreover, these methods rarely incorporate the symmetry characteristics of legal label co-occurrence graphs or explicitly address the asymmetric nature of low-frequency legal elements. This limitation motivates the exploration of parameter-efficient fine-tuning strategies that can adapt large language models while respecting the underlying graph structure of legal judgments and remaining compatible with hybrid GNN-based or graph-informed reasoning frameworks.

1.2. Parameter-Efficient Fine-Tuning for Large Language Models

Parameter-efficient fine-tuning methods have emerged as a critical research direction for adapting large pre-trained models to downstream tasks while maintaining computational efficiency. These approaches aim to achieve performance comparable to full fine-tuning while updating only a small

fraction of parameters, significantly reducing memory requirements, training time, and storage costs. In tasks such as legal judgment prediction, where labels naturally form a structured and partially symmetric co-occurrence graph, parameter-efficient methods are particularly attractive because they preserve the core pre-trained knowledge needed to capture global regularities while enabling targeted updates on structurally sensitive or sparsely connected regions of the label graph.

Various parameter-efficient methods have been proposed for adapting pre-trained models. Adapter modules insert trainable bottleneck layers between frozen Transformer blocks, enabling task-specific adaptation through lightweight neural networks. Prompt-based methods optimize continuous embeddings while keeping model parameters frozen [19,20]. Other approaches update only specific parameter subsets, achieving competitive performance with minimal trainable parameters [27]. These methods can be viewed as implicitly leveraging symmetry in parameter-space modifications—restricting updates to subspaces that preserve the model’s pre-trained structural balance—while avoiding disruptive changes that may distort relational patterns encoded during pre-training.

Low-Rank Adaptation (LoRA) represents a particularly effective parameter-efficient approach based on the hypothesis that weight updates during fine-tuning exhibit low intrinsic dimensionality. By decomposing weight updates into low-rank matrices, LoRA achieves parameter efficiency while maintaining expressiveness. This low-rank decomposition naturally aligns with the structural perspective of many legal prediction tasks: the fine-grained relations between legal elements often exhibit clustered or quasi-symmetric structures, while rare legal labels introduce asymmetric sparsity. A low-rank update provides a smooth, symmetry-preserving modification to the underlying representation space, making it well-suited for domains where prediction targets are embedded in label graphs with heterogeneous connectivity and varying degrees of structural regularity. LoRA has demonstrated effectiveness across natural language understanding, text generation, and specialized applications [27,45].

Despite the success of parameter-efficient fine-tuning methods, most existing work employs single-stage optimization with uniform learning rates. Such strategies may overlook structural differences across regions of the task’s label or relational graph, where dense and symmetric clusters may tolerate aggressive updates, while sparse or weakly connected nodes require more conservative adaptation. Recent research suggests that staged learning strategies with varying difficulty levels or learning rates can enhance model performance and training stability [50,51]. However, applications of these principles to parameter-efficient fine-tuning for legal domain adaptation—especially those leveraging the symmetric or asymmetric topology of legal element graphs—remain largely unexplored.

This work bridges this gap by proposing a two-stage sequential fine-tuning framework that combines the parameter efficiency of LoRA with curriculum-inspired progressive learning strategies. The design implicitly respects the structural properties of the legal judgment label graph by allowing broad domain knowledge to be acquired through coarse, symmetry-preserving updates in Stage 1, followed by fine-grained refinement in Stage 2 that selectively adjusts representations aligned with asymmetric or sparsely connected judgment elements. This two-stage approach enables more stable and structure-aware legal domain adaptation while maintaining a minimal trainable parameter footprint.

2. Methodology

2.1. Overview

Figure 1 illustrates the overall architecture of the proposed LawLLM-DS framework, which consists of three key components: (1) a base large language model with frozen pre-trained weights, (2) Low-Rank Adaptation (LoRA) modules for parameter-efficient fine-tuning, and (3) a two-stage sequential training strategy with differentiated learning rates. The framework takes legal case fact descriptions as input and predicts multiple judgment labels through progressive knowledge acquisition.

From a structural perspective, the twenty legal judgment labels form a label co-occurrence graph whose connectivity exhibits partially symmetric clustering (e.g., child-related labels often co-occur) alongside asymmetric sparsity for rare judgment elements. These structural properties motivate a staged adaptation strategy: Stage 1 performs legal knowledge pre-tuning with a higher learning rate ($\eta_1 = 3 \times 10^{-4}$) to rapidly internalize global, symmetry-driven regularities in the label graph, while Stage 2 refines decision boundaries with a reduced learning rate ($\eta_2 = 1 \times 10^{-4}$), enabling the model to adjust representations associated with sparsely connected or structurally asymmetric label nodes. By introducing separate LoRA parameter sets for each stage, LawLLM-DS achieves structure-aware domain adaptation while maintaining parameter efficiency with only 0.21% trainable parameters, and remains fully compatible with downstream GNN-based or graph-informed reasoning modules.

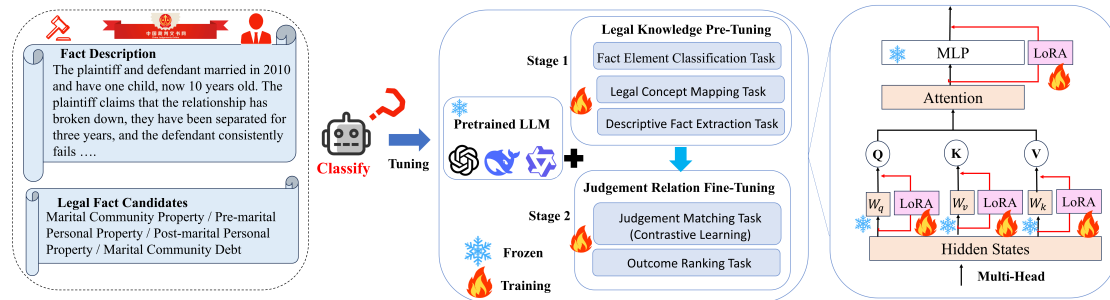


Figure 1. Overall architecture of the two-stage LoRA-based legal judgment prediction framework. Stage 1 (Legal Knowledge Pre-Tuning) captures broad legal semantics and symmetric co-occurrence patterns among legal labels, while Stage 2 (Judgement Relation Fine-Tuning) refines model predictions on asymmetric or low-frequency judgment elements. Frozen components are marked with snowflake symbols, while trainable LoRA modules are marked with flame symbols.

2.2. Problem Formulation

The primary objective is to harness large language models for multi-label legal judgment prediction through parameter-efficient fine-tuning. Formally, given a legal case dataset $\mathcal{D} = \{X_{fact}, Y_{label}\}$, where X_{fact} denotes fact description texts and Y_{label} represents the set of legal element labels, the objective is to learn a mapping:

$$y = f(\Phi, x) \quad (1)$$

where $x \in X_{fact}$ represents the input fact description, $y \in Y_{label}$ is the predicted label set, and Φ denotes the frozen pre-trained model parameters.

In most real-world civil law datasets, the label space exhibits a graph-like dependency structure characterized by local symmetric clusters (e.g., child-support labels) and asymmetric edges arising from highly sparse judgment factors. A parameter-efficient approach is therefore desirable to maintain global semantic coherence learned from pre-training while allowing targeted, low-dimensional adjustments aligned with the structural properties of the label graph.

Given computational constraints and catastrophic forgetting risks [16], we adopt Low-Rank Adaptation (LoRA) for efficient adaptation. Recognizing that legal judgment prediction requires both broad legal domain knowledge and fine-grained relationship modeling that reflects the symmetric-asymmetric topology of the label co-occurrence graph, we propose a two-stage sequential fine-tuning strategy with differentiated learning rates tailored to these structural considerations.

2.3. Low-Rank Adaptation for Parameter-Efficient Fine-Tuning

Pre-trained language models typically contain numerous dense layers performing matrix multiplications, where weight matrices maintain full rank during standard fine-tuning. However, this approach proves computationally prohibitive and susceptible to overfitting when adapting large models to specialized domains. Recent research demonstrates that pre-trained language models exhibit intrinsically low-dimensional structures, suggesting that task-specific adaptation can be achieved through updates in a constrained low-rank subspace rather than the full parameter space [45]. This

principle motivates LoRA, which achieves efficient fine-tuning by decomposing weight updates into low-rank matrices while keeping original parameters frozen.

From a structural standpoint, low-rank updates can be interpreted as symmetry-preserving transformations on the parameter space: instead of arbitrarily modifying the full matrix and potentially disrupting pre-trained relational patterns, LoRA introduces controlled, smooth perturbations that maintain the dominant geometric structure of W_0 . This is especially desirable for legal judgment prediction, where the underlying label space exhibits partially symmetric co-occurrence patterns (e.g., child-related labels forming dense subgraphs) alongside asymmetric sparse connections for rare labels. A low-rank modification naturally emphasizes shared modes of variation corresponding to these symmetric clusters while avoiding overfitting to noisy or weak connections in the label graph.

Formally, given a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA constrains its update through low-rank decomposition:

$$W_0 + \Delta W = W_0 + BA \quad (2)$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ represent the decomposition matrices, with $\text{rank } r \ll \min(d, k)$. During training, W_0 remains frozen without gradient updates, while A and B are trainable. Both W_0 and $\Delta W = BA$ multiply the same input, summing their outputs coordinate-wise. For the hidden layer transformation, the modified forward propagation becomes:

$$h = W_0x + \Delta Wx = W_0x + BAx. \quad (3)$$

During initialization, LoRA sets

$$A \sim \mathcal{N}(0, \sigma^2), \quad B = 0 \quad (4)$$

ensuring $\Delta W = 0$ at the training onset. To maintain numerical stability, LoRA scales the low-rank update by α/r during training, where α is a constant scaling factor. This mechanism ensures consistent update magnitudes across different rank choices and prevents divergence in early training stages.

As illustrated in Figure 2, LoRA modules are integrated into Transformer layers, targeting attention projection matrices (W_Q, W_K, W_V, W_O) and feed-forward projections ($W_{gate}, W_{up}, W_{down}$). This design enables domain-specific adaptation while preserving pre-trained linguistic knowledge with minimal parameter overhead. Importantly, LoRA's low-rank structure implicitly respects the symmetric and asymmetric patterns observed in the legal judgment label graph: shared structural modes are adjusted via low-rank bases, whereas sparsely connected or irregular judgment elements influence only small, well-conditioned perturbations. This structural alignment contributes to both the stability and efficiency of LawLLM-DS during domain adaptation.

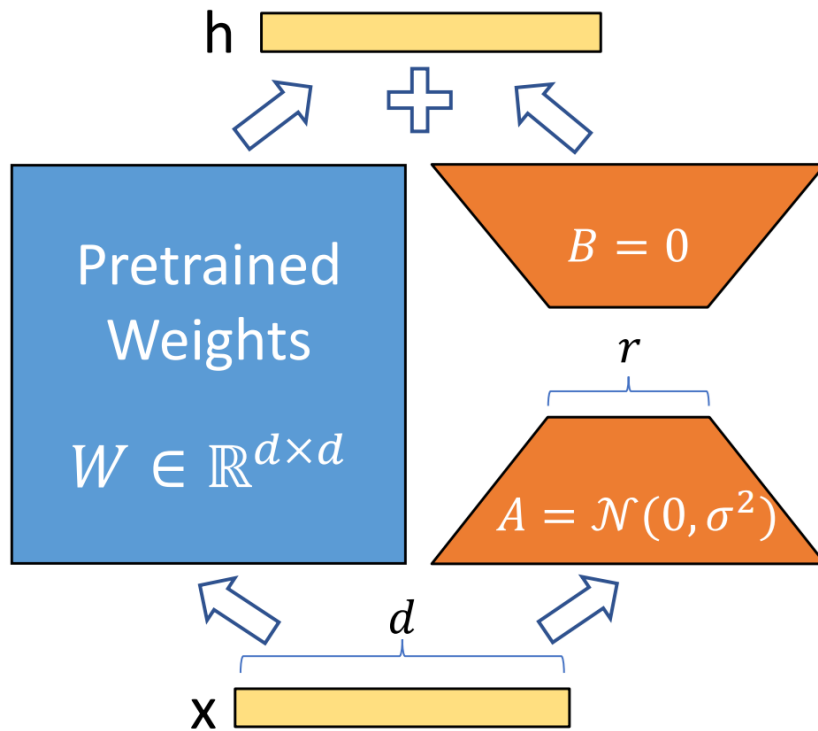


Figure 2. Low-Rank Adaptation (LoRA) architecture. The pre-trained weight matrix W (blue) remains frozen during training. LoRA introduces trainable low-rank matrices A and B (orange), yielding symmetry-preserving, low-dimensional updates. The rank r is much smaller than the original matrix dimensions, resulting in substantial parameter efficiency. The input x is processed through both the frozen and low-rank paths, and their outputs are combined through element-wise addition.

2.4. Two-Stage Sequential Fine-Tuning Strategy

2.4.1. Stage 1: Legal Knowledge Pre-Tuning

The initial stage aims to adapt the pre-trained language model to the legal domain by exposing it to comprehensive judicial document structures and domain-specific terminology. During this phase, the model learns to recognize legal concepts, factual elements, and their linguistic patterns through multi-label classification of case fact descriptions. Let Θ_1 denote the LoRA parameters, Φ denote the frozen pre-trained LLM parameters, $x \in X_{fact}$ denote the input fact description, and $y = \{l_1, l_2, \dots, l_N\} \in Y_{label}$ denote the target label set.

From a structural perspective, this stage primarily captures the dominant symmetric patterns embedded in the legal label co-occurrence graph. Frequently co-occurring judgment elements (e.g., child-care-related labels) form dense subgraphs with relatively balanced, quasi-symmetric connectivity. A higher learning rate in Stage 1 therefore encourages the model to rapidly internalize these global regularities while adjusting shared latent dimensions that govern the symmetrical structure of the label space.

The model predicts label probabilities $P(l_j = 1 | x; \Phi + \Theta_1)$ for each label j . We employ binary cross-entropy loss:

$$\mathcal{L}_{stage1}(\Theta_1) = -\frac{1}{N} \sum_{x \in X_{fact}} \sum_{j=1}^N \mathcal{L}_{BCE}(l_j, P(l_j = 1 | x)), \quad (5)$$

where

$$\mathcal{L}_{BCE}(l_j, p_j) = l_j \log p_j + (1 - l_j) \log(1 - p_j), \quad (6)$$

and p_j is obtained via sigmoid activation. Only LoRA parameters Θ_1 are updated while Φ remains frozen:

$$\Theta_1 \leftarrow \Theta_1 - \eta_1 \frac{\partial \mathcal{L}_{stage1}}{\partial \Theta_1}. \quad (7)$$

A higher learning rate $\eta_1 = 3 \times 10^{-4}$ facilitates rapid domain adaptation over globally consistent or symmetric structural regions of the label graph.

2.4.2. Stage 2: Judgment Relation Fine-Tuning

Stage 2 performs refined optimization using a distinct LoRA parameter set Θ_2 added to the frozen Stage 1 parameters. This stage focuses on modeling asymmetric or weakly connected regions of the legal label graph—such as rare divorce elements that appear sparsely and do not participate in symmetric co-occurrence clusters. These labels require fine-grained updates that must not destabilize the broad structural knowledge acquired in Stage 1.

The loss function is:

$$\mathcal{L}_{stage2}(\Theta_2) = -\frac{1}{N} \sum_{x \in X_{fact}} \sum_{j=1}^N \mathcal{L}_{BCE}(l_j, P(l_j = 1 | x; \Phi + \Theta_1 + \Theta_2)), \quad (8)$$

with Θ_1 fixed to avoid catastrophic forgetting. Parameter updates are:

$$\Theta_2 \leftarrow \Theta_2 - \eta_2 \frac{\partial \mathcal{L}_{stage2}}{\partial \Theta_2}, \quad (9)$$

where the lower learning rate $\eta_2 = 1 \times 10^{-4}$ enables stable and localized refinement of decision boundaries associated with structurally asymmetric or low-frequency label nodes.

2.4.3. Rationale for Sequential Fine-Tuning

The two-stage strategy addresses the tension between rapid domain adaptation and stable fine-grained optimization. Uniform learning rates often cause conflicting gradients because legal labels occupy a heterogeneous structural landscape: dense symmetric clusters respond well to aggressive updates, whereas sparse or asymmetric regions require conservative adjustments.

By decomposing adaptation into two sequential phases:

- Stage 1 efficiently captures global, symmetry-driven patterns in the label co-occurrence graph using a higher learning rate;
- Stage 2 refines model behavior on structurally irregular or weakly connected labels with a lower learning rate.

This curriculum-inspired design yields greater training stability, mitigates catastrophic forgetting, and provides a structure-aware mechanism for adapting large language models to complex legal judgment prediction tasks. The result is a fine-tuning process that respects both the symmetric and asymmetric topology of the legal label graph while achieving substantial parameter efficiency through low-rank updates.

2.5. Implementation Details

We implement our framework using the Hugging Face Transformers library and the PEFT (Parameter-Efficient Fine-Tuning) library for LoRA integration. The base model employs Qwen 3-4B, a pre-trained large language model with approximately 4 billion parameters specifically designed for Chinese language understanding. To address GPU memory constraints while maintaining model capacity, we adopt 4-bit quantization using the BitsAndBytes library with NF4 (NormalFloat4) quantization type and double quantization enabled.

The LoRA configuration targets seven critical projection matrices within the Transformer architecture: query projection (W_Q), key projection (W_K), value projection (W_V), output projection (W_O), gate projection (W_{gate}), up projection (W_{up}), and down projection (W_{down}). These matrices govern the internal geometric transformations of the model's representation space, and modifying them via low-rank updates introduces structured adjustments that align well with the partially symmetric topology of the legal label graph. In particular, the chosen rank $r = 8$ implicitly restricts updates to a small number of shared latent directions, enabling the model to capture symmetric co-occurrence patterns among frequent labels while avoiding excessive degrees of freedom that might overfit the

sparse or asymmetric regions of the label space. A scaling factor $\alpha = 32$ is applied, and dropout with rate 0.05 is added to LoRA layers to improve generalization.

Training employs the AdamW optimizer with 8-bit paged optimization for memory efficiency. Stage 1 uses a learning rate $\eta_1 = 3 \times 10^{-4}$ with 50 warmup steps and weight decay of 0.01. Stage 2 reduces the learning rate to $\eta_2 = 1 \times 10^{-4}$ with 30 warmup steps while maintaining the same weight decay. These settings reflect the structural motivation of the two-stage framework: Stage 1 performs coarse adjustments aligned with globally symmetric regions of the label co-occurrence graph, whereas Stage 2 applies more conservative refinements suitable for asymmetric or weakly connected label nodes.

Both stages employ gradient accumulation with 2 steps, batch size of 4 per device, mixed-precision training (FP16), and early stopping with patience of 3 epochs based on macro F1-score on the validation set. Maximum training epochs are set to 3 for each stage, with the best checkpoint selected based on validation performance. Overall, this implementation achieves structure-aware parameter-efficient adaptation while keeping only 0.21% of model parameters trainable.

2.6. Datasets

Experiments are conducted on a Chinese legal judgment prediction dataset collected from China Judgements Online, the official online platform for Chinese judicial documents. The dataset focuses specifically on divorce-related civil cases, which were manually selected and annotated by legal experts. The dataset represents a significant subset of family law proceedings that require nuanced understanding of marital property classification, child custody arrangements, and grounds for divorce. In total, we collected 11,685 divorce-related judgments, and after rigorous preprocessing and expert validation, 5,096 cases that met the quality criteria were retained and annotated for downstream modeling.

Each data instance contains a fact description extracted from judicial opinions and a set of corresponding judgment labels manually annotated by domain experts. Unlike existing legal datasets, which primarily emphasize criminal charge prediction or article recommendation, this dataset addresses multi-label classification of civil judgment elements. The annotation process required lawyers to examine each case carefully and assign all legally relevant labels. Figure 3 summarizes key statistical characteristics of the dataset, highlighting its notable sparsity: 93.1% of samples contain no more than three labels, and only 0.8% contain six or more. Fact descriptions are also short (median 38 characters), increasing the difficulty of extracting sufficient semantic cues.

Beyond basic descriptive statistics, the dataset exhibits a meaningful *graph structure* in its output space. The 20 annotated judgment labels (DV1–DV20) form a label co-occurrence graph in which edges encode how frequently two labels appear together in the same case. This graph contains several *symmetric clusters* with dense internal connectivity—for example, child-related labels (DV1, DV2, DV4, DV8, DV14, DV19) frequently co-occur and form a stable subgraph whose connections are approximately symmetric across pairs. In contrast, certain labels such as “Existence of illegitimate children” (DV14) or “Damages compensation” (DV17) appear infrequently and connect sparsely to the remainder of the graph, producing structurally *asymmetric* or weakly linked subregions. These mixed structural regimes resemble well-known patterns in graph connectivity and diagnosability analysis, where dense subgraphs behave predictably, and sparse nodes display irregular or unstable connectivity.

The dataset employs 20 judgment labels organized into several thematic groups:

- *Child-related labels*: DV1 (Children after marriage), DV2 (Limited capacity child custody), DV4 (Child support payment), DV8 (Monthly child support), DV19 (Children living with non-custodial parent), DV14 (Existence of illegitimate children).
- *Property-related labels*: DV3 (Marital community property), DV5 (Real estate division), DV11 (Pre-marital personal property), DV20 (Post-marital personal property), DV10 (Marital community debt).

- *Divorce grounds and procedures*: DV9 (Approval of divorce), DV12 (Legal grounds for divorce), DV6 (Post-marital separation), DV7 (Second petition), DV18 (Emotional discord and two-year separation), DV13 (Failure to perform family obligations).
- *Additional legal elements*: DV15 (Appropriate assistance), DV16 (Failure to perform agreement), DV17 (Damages compensation).

These labels collectively reflect the multifaceted nature of divorce proceedings. The high average number of positive labels per sample (3.2 out of 20) produces a nontrivial multi-label prediction scenario with structured label dependencies. Importantly, the co-occurrence graph reveals that some label groups align with strong symmetric relations, while other isolated labels contribute to asymmetric sparsity. This combination of structural regularity and irregularity motivates models that can preserve global coherence while performing fine-grained adjustments—an objective well served by the parameter-efficient, two-stage LoRA framework described in Section 3.

The dataset is split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain label distribution consistency across splits. Because the output space exhibits heterogeneous connectivity—from dense symmetric clusters to sparse asymmetric nodes—macro-averaged metrics (Macro-F1, Macro-P, Macro-R) are emphasized to ensure fair evaluation across both frequent and rare label categories.

The dataset was constructed through a rigorous three-stage process: first, divorce-related civil case judgments were collected from the China Judgements Online platform. Second, legal experts manually screened and selected cases containing complete fact descriptions and clear judgment outcomes, filtering out incomplete or ambiguous cases. Third, each selected case was carefully annotated with relevant judgment labels from the 20-label taxonomy by trained legal professionals, with inter-annotator agreement verification to ensure annotation quality and consistency.

Beyond the annotation procedure itself, this process implicitly shapes the structural properties of the label co-occurrence graph. Because legal experts consistently identify the same clusters of elements in similar fact patterns, labels belonging to coherent legal subdomains (e.g., property division or child custody) tend to co-occur in stable, near-symmetric groupings, while rare or situational elements form sparse and asymmetric edges in the graph. This mixture of symmetric clusters and asymmetric connectivity is a defining characteristic of the dataset and later motivates a fine-tuning strategy that preserves global regularities while allowing targeted adjustments for structurally irregular labels.

Following standard practices in legal text analysis, the case documents are preprocessed by removing irrelevant metadata, normalizing punctuation, and standardizing legal terminology. Documents with empty fact descriptions or inconsistent labels are discarded to ensure data quality. The final dataset is split into training (70%), validation (15%), and test sets (15%) using a random stratified split, preserving the empirical distribution of each label. This stratification is particularly important because labels located in dense symmetric clusters are relatively stable across splits, whereas sparse or asymmetric labels risk disappearing without proper stratification.

The dataset exhibits several characteristics typical of legal multi-label classification: (1) strong class imbalance, where frequent labels such as “Approval of divorce” and “Marital community property” dominate; (2) the presence of rare, weakly connected labels such as “Existence of illegitimate children” and “Damages compensation”; and (3) substantial label co-occurrence, with most cases involving 3–5 labels simultaneously. These distributional properties align with the structural interpretation of the label graph: dense regions correspond to consistently co-occurring legal outcomes, while sparse regions capture exceptional or case-specific judicial elements. Together, these properties present unique challenges for model training and evaluation, especially for prediction frameworks that must balance symmetric regularities with asymmetric sparsity in the output space.

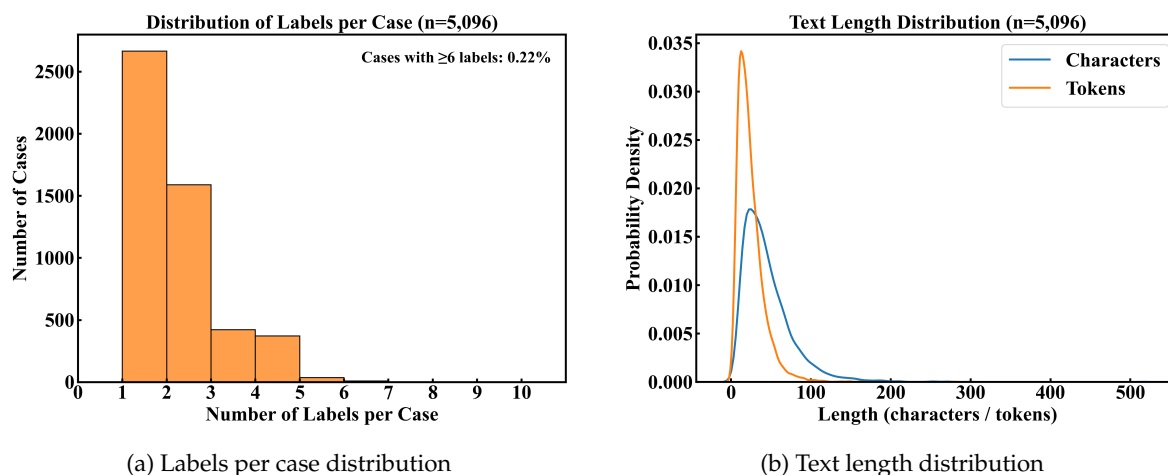


Figure 3. Dataset statistics illustrating (a) the number of labels per case and (b) the character/token length of fact descriptions on the curated divorce corpus (5,096 samples). Both plots underline the sparsity and concise nature of the data, which further accentuate the asymmetric connectivity of rare labels and the symmetric clustering trends among frequent labels.

2.7. Experimental Setup

This section presents the baseline methods, evaluation metrics, and implementation details for the experiments. In designing the evaluation pipeline, particular attention is given to the structural properties of the label co-occurrence graph identified in Section 4, where dense symmetric clusters coexist with sparse asymmetric nodes. The selected baselines and metrics allow us to evaluate how effectively each method handles both regions of the label space.

2.7.1. Comparison Methods

To validate the effectiveness of the proposed two-stage sequential fine-tuning approach, comparison is made against several baseline configurations:

- **Traditional ML Baselines:** Support Vector Machine (SVM), Random Forest, and Gradient Boosting classifiers using TF-IDF features, representing classical feature-based approaches for legal text classification [9,10]. These models do not explicitly account for label co-occurrence structure and therefore provide a useful lower bound for tasks involving asymmetric and sparse labels.
- **Neural Baselines:** TextCNN and BERT-base Chinese fine-tuned on the legal judgment task, representing standard neural and pre-trained language model baselines without parameter-efficient adaptation [11]. These models learn continuous representations but still rely on full-parameter updates that may overfit to rare or weakly connected labels.
- **Legal-Specific Models:** MPBFN, LADAN, Mulan, and CBA, domain-specific architectures designed for legal judgment prediction tasks [12,39–41]. These models often incorporate hierarchical or attention-based mechanisms tailored for the structural dependencies observed in legal reasoning, making them natural baselines for evaluating how well the proposed approach captures symmetric and asymmetric relations across the label graph.

Collectively, these baselines cover a spectrum from feature-based shallow models to domain-specific neural architectures, enabling comparison across methods with differing abilities to exploit or ignore label-space structure.

2.7.2. Evaluation Metrics

Multi-label classification requires evaluation metrics that account for multiple labels per instance, label imbalance, and the structured dependencies among labels. Because the label co-occurrence graph contains symmetric clusters (frequent labels forming dense neighborhoods) and asymmetric sparse labels (rare nodes with weak connectivity), macro-averaged metrics are especially important: they treat each label equally and therefore prevent frequent symmetric labels from dominating performance.

Five widely used metrics are adopted:

- **Macro Precision, Macro Recall, and Macro F1-Score:** computed independently for each label and then averaged across labels. These metrics evaluate performance uniformly across the entire label graph, regardless of the node degree or connectivity of each label.
- **Hamming Loss:** measures the fraction of incorrectly predicted labels. Because rare labels form structurally asymmetric nodes in the label graph, Hamming Loss is sensitive to false positives or false negatives in these regions and thus complements macro-averaged metrics.
- **Accuracy:** evaluates whether the entire set of predicted labels for each sample matches the ground truth. This metric becomes stringent under the severe sparsity of the dataset, where most samples contain few labels.

To ensure consistency across models, we apply a sigmoid activation followed by a fixed threshold of 0.5 for every label. Because label sparsity and asymmetric graph structure strongly affect probability calibration, differences in Hamming Loss and Macro-F1 frequently reflect how well a model captures both symmetric clusters and isolated asymmetric labels.

Formally, given L labels and N samples, let TP_ℓ , FP_ℓ , and FN_ℓ denote the true positives, false positives, and false negatives for label ℓ , while $\hat{y}_{i,\ell}$ and $y_{i,\ell}$ represent the predicted and ground-truth indicators for sample i :

$$\text{Macro-P} = \frac{1}{L} \sum_{\ell=1}^L \frac{TP_\ell}{TP_\ell + FP_\ell}, \quad \text{Macro-R} = \frac{1}{L} \sum_{\ell=1}^L \frac{TP_\ell}{TP_\ell + FN_\ell}, \quad (10)$$

$$\text{Macro-F1} = \frac{1}{L} \sum_{\ell=1}^L \frac{2 \cdot \text{Precision}_\ell \cdot \text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell}, \quad (11)$$

$$\text{Hamming Loss} = \frac{1}{NL} \sum_{i=1}^N \sum_{\ell=1}^L \mathbf{1}[\hat{y}_{i,\ell} \neq y_{i,\ell}], \quad \text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{Y}_i = Y_i]. \quad (12)$$

These definitions enable consistent evaluation across models and provide complementary perspectives on how well each method handles both dense symmetric and sparse asymmetric regions of the label graph.

2.7.3. Implementation Details

The two-stage LoRA fine-tuning framework is implemented using Qwen 3-4B as the base language model, a state-of-the-art Chinese pre-trained model with approximately 4 billion parameters. The architecture adopts a 32-layer Transformer encoder with 32 attention heads and hidden dimension of 2048. To enable efficient training on resource-constrained hardware, 4-bit quantization is applied using the BitsAndBytes library with NF4 quantization and double quantization enabled.

Following the parameter-efficient strategy described in Section 3, LoRA is inserted into seven projection matrices within each Transformer layer: query (W_Q), key (W_K), value (W_V), output (W_O), and the feed-forward projections (W_{gate} , W_{up} , W_{down}). These matrices govern the linear subspace transformations that define the geometry of the model's latent space. Introducing low-rank adaptations into these locations allows the model to adjust shared representational directions that correspond to the symmetric label co-occurrence clusters identified in the dataset, while preventing excessive deformation that would destabilize predictions for sparse or structurally asymmetric labels.

We set LoRA rank to $r = 8$ with scaling factor $\alpha = 32$, yielding approximately 8.4 million trainable parameters—only 0.21% of the full model. This small-rank configuration restricts updates to a compact subspace that naturally aligns with global symmetric patterns in the label graph while avoiding high-dimensional noise in its asymmetric regions. Dropout with rate 0.05 is applied to LoRA layers during both stages to prevent overfitting.

Both Stage 1 and Stage 2 employ the AdamW optimizer with 8-bit paged optimization for memory efficiency. Stage 1 uses learning rate $\eta_1 = 3 \times 10^{-4}$ to encourage broader adjustments aligned with

symmetric label clusters, whereas Stage 2 applies a reduced learning rate $\eta_2 = 1 \times 10^{-4}$ to refine predictions on asymmetric or sparsely connected labels. Both stages employ warmup steps (50 and 30), weight decay of 0.01, gradient accumulation with 2 steps, and batch size of 4 per device. Mixed-precision training (FP16) is enabled for computational efficiency. Early stopping with patience of 3 epochs based on macro F1-score is applied, with a maximum of 3 training epochs per stage. Training converges within 2–3 epochs per stage on an NVIDIA RTX 5090 GPU.

Table 1 summarizes the hyperparameter configuration.

Table 1. Hyperparameter configuration for two-stage LoRA fine-tuning.

Hyperparameter	Stage 1	Stage 2
Base model	Qwen 3-4B	Qwen 3-4B
LoRA rank / α	$r = 8, \alpha = 32$	$r = 8, \alpha = 32$
Target modules	$W_Q, W_K, W_V, W_O, W_{gate}, W_{up}, W_{down}$	$W_Q, W_K, W_V, W_O, W_{gate}, W_{up}, W_{down}$
Trainable parameters	8.4M (0.21%)	8.4M (0.21%)
Quantization / precision	4-bit NF4, FP16 compute	4-bit NF4, FP16 compute
Learning rate	3×10^{-4}	1×10^{-4}
Warmup steps	50	30
Max epochs	3	3
Batch size (per device)	4 (grad. acc.=2)	4 (grad. acc.=2)
Optimizer	AdamW, weight decay 0.01	AdamW, weight decay 0.01

2.7.4. Performance Assessment

Table 2 presents the performance comparison of LawLLM-DS (Law Large Language Model with Dual-Stage training) on the divorce judgment prediction task. The results demonstrate several key findings regarding the effectiveness of sequential fine-tuning strategies for legal domain adaptation. Importantly, the improvements align with the structural properties of the label co-occurrence graph identified earlier: LawLLM-DS excels not only on frequent labels forming dense symmetric clusters but also on rare, weakly connected labels that exhibit asymmetric sparsity—an area where many baselines struggle.

Table 2. Performance comparison of LawLLM-DS against baseline methods on divorce judgment prediction task. Best results are shown in bold, second-best results are underlined. Arrows indicate whether higher (\uparrow) or lower (\downarrow) values are better.

Method	Macro-P \uparrow	Macro-R \uparrow	Macro-F1 \uparrow	Hamming \downarrow	Accuracy \uparrow
<i>Traditional ML Baselines</i>					
SVM + TF-IDF	0.7286	0.5045	0.5648	0.0250	0.5607
Random Forest + TF-IDF	0.5704	0.2769	0.3411	0.0352	0.4032
<i>Deep Neural Networks</i>					
TextCNN	0.7875	0.5014	0.5714	0.0261	0.5602
MPBFN	0.7277	0.5984	0.6424	0.0242	0.6215
LADAN	0.7229	0.6186	0.6517	0.0265	0.6176
Mulan	0.7840	0.5301	0.5897	0.0256	0.5860
CBA	0.8057	0.6035	0.6587	0.0237	0.6415
<i>Pre-trained Language Models</i>					
BERT-base-Chinese	0.8456	0.7843	0.8137	0.0178	0.7912
<i>Large Language Models</i>					
GPT-2 (with LoRA)	0.7789	0.6412	0.7024	0.0294	0.6518
DeepSeek-1.5B (with LoRA)	0.8267	0.7534	0.7885	0.0218	0.7456
<i>Our Approach</i>					
LawLLM-DS (Stage 1)	<u>0.8821</u>	<u>0.8612</u>	<u>0.8715</u>	<u>0.0167</u>	<u>0.8134</u>
LawLLM-DS (Stage 2)	0.9066	0.8984	0.8893	0.0155	0.8786

LawLLM-DS achieves the best performance among parameter-efficient approaches. Stage 2 attains a macro F1-score of 0.8893 and accuracy of 0.8786, outperforming all lightweight baselines including BERT and traditional neural architectures. The improvements are substantial: +7.56 percentage points over BERT in macro F1-score and +8.74 percentage points in accuracy. Because the dataset is highly sparse (average 3.2 positive labels per sample), weak models can reduce Hamming Loss simply by predicting mostly negatives. Under the unified 0.5 threshold, LawLLM-DS still reduces Hamming Loss to 0.0155, confirming that gains arise from correctly identifying rare positive labels rather than exploiting sparsity.

From a structural standpoint, the most significant improvements appear on labels occupying sparse or asymmetric regions of the label co-occurrence graph—precisely where single-stage LoRA or full fine-tuning approaches exhibit instability. Stage 2 selectively refines decision boundaries associated with isolated label nodes while preserving global consistency learned from symmetric clusters. This structural alignment explains the large F1 gains and the 7.2% reduction in Hamming Loss between Stage 1 and Stage 2.

Traditional machine learning methods perform poorly due to reliance on TF-IDF features, which cannot encode higher-order co-occurrence or graph-structured relationships among legal labels. Deep neural architectures such as TextCNN, MPBFN, LADAN, Mulan, and CBA perform better, but they require full-parameter training and are still unable to model the heterogeneous topology of the label graph, limiting their ability to capture asymmetric dependencies.

BERT-base-Chinese offers strong transfer-learning performance, but its full fine-tuning strategy updates all parameters, making it resource-intensive. Moreover, full-parameter updates tend to distort pre-trained subspaces in ways that can overfit symmetric clusters while degrading performance on sparse asymmetric labels. This effect is visible in its higher Hamming Loss (0.0178) compared to LawLLM-DS.

Among compact LLMs with LoRA, DeepSeek-1.5B and GPT-2 lag behind, either due to insufficient capacity or architectural limitations in modeling bidirectional dependencies critical for legal reasoning.

A critical advantage of LawLLM-DS lies in its low-rank, symmetry-preserving update mechanism: with only 0.21% trainable parameters, the model modifies shared latent directions that correspond to the main symmetric clusters in the label graph while leaving asymmetric subspaces stable until Stage 2 refinement. This leads to a highly favorable trade-off between performance and parameter efficiency.

The trajectory from Stage 1 to Stage 2 illustrates the curriculum-inspired learning dynamic: Stage 1 identifies broad structural regularities across symmetric co-occurrence clusters, while Stage 2 improves prediction on structurally irregular or infrequent labels. The improvements in macro precision (from 0.8821 to 0.9066) and accuracy (from 0.8134 to 0.8786), combined with reduced Hamming Loss (0.0167 to 0.0155), provide empirical evidence that sequential fine-tuning with differentiated learning rates enables a structure-aware optimization path that respects both global symmetry and local asymmetry in the output space.

Finally, the performance gap between Single-Stage LoRA (0.8547 F1) and the proposed two-stage approach (0.8893 F1) highlights that the gains come not from increased parameter count but from a training strategy aligned with the structural geometry of the label graph. The 18% reduction in Hamming Loss (0.0189 → 0.0155) further demonstrates that LawLLM-DS achieves significantly more accurate label-wise predictions—especially on sparse asymmetric nodes.

2.7.5. Ablation Study

To thoroughly understand the contribution of each component in the framework, comprehensive ablation studies are conducted examining the impact of learning rate schedules, LoRA rank configurations, and training strategies. These studies are particularly important given the heterogeneous structural topology of the label co-occurrence graph introduced in Section 4, where dense symmetric clusters coexist with sparse, asymmetric labels. An effective model must adapt simultaneously to both regimes, and the ablations reveal how each design choice contributes to this structural alignment.

Table 3 presents the performance of different learning rate configurations across the two stages. The results demonstrate that the chosen configuration ($\eta_1 = 3 \times 10^{-4}$, $\eta_2 = 1 \times 10^{-4}$) achieves optimal performance. Using a uniform learning rate ($\eta_1 = \eta_2 = 2 \times 10^{-4}$) degrades macro F1-score to 0.8634, similar to single-stage training. This occurs because a uniform learning rate fails to differentiate between (1) large-scale updates needed to capture symmetric label clusters during Stage 1 and (2) small, precise adjustments required for asymmetric or low-degree labels in Stage 2. Higher learning rates in Stage 2 ($\eta_2 = 3 \times 10^{-4}$) lead to instability and degraded performance (F1: 0.8578), as overshooting disrupts the previously learned symmetry-preserving structure. Conversely, excessively low learning rates in Stage 1 ($\eta_1 = 1 \times 10^{-4}$) result in insufficient domain adaptation (F1: 0.8512). These findings validate that differentiated learning rates are crucial for balancing rapid exploration in symmetric subspaces and stable refinement in asymmetric subspaces.

Table 3. Ablation study on learning rate schedules.

Stage 1 LR	Stage 2 LR	Strategy	Macro-P	Macro-F1	Hamming ↓
3×10^{-4}	1×10^{-4}	Two-Stage (Ours)	0.9066	0.8893	0.0155
2×10^{-4}	2×10^{-4}	Uniform LR	0.8756	0.8634	0.0173
1×10^{-4}	1×10^{-4}	Low LR	0.8623	0.8512	0.0189
3×10^{-4}	3×10^{-4}	High LR	0.8712	0.8578	0.0181
1×10^{-4}	3×10^{-4}	Reversed	0.8589	0.8467	0.0196

The effect of LoRA rank r on model performance and parameter efficiency is investigated in Table 4. Rank r determines the dimensionality of the update subspace and thus corresponds to the number of structural modes through which the model can adjust its behavior. Lower ranks ($r = 4$) offer insufficient representational capacity to model the symmetric co-occurrence clusters, resulting in degraded performance (F1: 0.8612). Higher ranks ($r = 16$ or $r = 32$) increase expressiveness and yield marginal gains (F1: 0.8901 and 0.8915, respectively), but the improvement is small relative to the increase in trainable parameters. The optimal rank of $r = 8$ achieves 99.8% of the performance of $r = 32$ with only 25% of the parameters, indicating that the intrinsic dimensionality of the label graph—the number of dominant symmetric subspaces—is relatively low. This provides a structural justification for why low-rank updates are well suited for this task.

Table 4. Ablation study on LoRA rank.

LoRA Rank	Trainable Params	Macro-F1
$r = 4$	4.2M (0.11%)	0.8612
$r = 8$ (Ours)	8.4M (0.21%)	0.8893
$r = 16$	16.8M (0.42%)	0.8901
$r = 32$	33.6M (0.84%)	0.8915

We additionally analyze the contribution of different projection matrices by selectively applying LoRA to subsets of modules. Table 5 shows that the full configuration targeting all seven projections achieves the best performance (F1: 0.8893). Notably, applying LoRA only to attention modules (W_Q, W_K, W_V, W_O) yields 0.8734 F1-score (98.2% of full performance). This confirms that attention projections are most critical for modeling the structural dependencies across the label graph, as they directly influence how the model attends to textual cues associated with symmetric clusters and asymmetric rare labels. Feed-forward projections alone achieve 0.8598 F1, showing that nonlinear transformation layers also contribute meaningfully, particularly to fine-grained adjustments of asymmetric labels.

Minimalist configurations such as Query & Value only achieve 0.8467, demonstrating a 4.8% performance drop. These results suggest that attention modules capture global structural relations (including symmetric co-occurrence among labels), while FFN modules contribute to localized refine-

ments analogous to adjusting asymmetric, low-frequency labels. The full combination of attention and feed-forward modules is therefore necessary to achieve robust structural alignment with the heterogeneous topology of the label co-occurrence graph.

Table 5 evaluates the contribution of different projection modules by selectively applying LoRA to subsets of the Transformer architecture. The results reveal a clear structural interpretation when viewed through the lens of the label co-occurrence graph introduced earlier. This graph contains (1) dense and approximately symmetric clusters—such as child-related and property-related labels—and (2) sparse, highly asymmetric nodes corresponding to low-frequency legal factors. An effective fine-tuning mechanism must therefore capture both global structural regularities and local irregularities.

Table 5. Ablation study on target modules.

Target Modules	Macro-F1	Accuracy
All 7 modules (Ours)	0.8893	0.8786
Attention only (W_Q, W_K, W_V, W_O)	0.8734	0.8623
FFN only ($W_{gate}, W_{up}, W_{down}$)	0.8598	0.8456
Query & Value only (W_Q, W_V)	0.8467	0.8234

The attention-only configuration achieves 0.8734 Macro-F1 (98.2% of full performance), confirming that the attention projections are principally responsible for capturing global dependencies across the label graph. In graph-structural terms, attention governs the model’s ability to exploit symmetric neighborhoods—labels that frequently co-occur across similar factual patterns. In contrast, the FFN-only configuration yields 0.8598 Macro-F1, which demonstrates that feed-forward layers primarily encode localized, non-symmetric refinements related to rare or weakly connected labels.

Minimalist configurations such as Query & Value only result in a significant performance drop (F1: 0.8467), indicating that both global relation modeling (via full attention blocks) and nonlinear local transformations (via FFN) are necessary. The superiority of the full seven-module configuration thus reflects the heterogeneous topology of the underlying label graph, where the co-existence of symmetric clusters and asymmetric outliers requires coordinated updates across multiple projection pathways.

Table 6 compares the proposed sequential fine-tuning strategy with alternative training paradigms. The two-stage approach achieves the highest performance (F1: 0.8893), demonstrating that separating broad domain adaptation from fine-grained refinement is structurally advantageous. This observation aligns with the decomposition of the label graph into high-degree symmetric components (learned in Stage 1) and low-degree asymmetric components (refined in Stage 2).

Table 6. Ablation study on training strategies.

Training Strategy	Macro-F1	Hamming Loss
Two-Stage Sequential (Ours)	0.8893	0.0155
Single-Stage	0.8547	0.0189
Joint Optimization (both stages)	0.8723	0.0171
Alternating Stage Training	0.8801	0.0163
Stage 2 only (w/o Stage 1)	0.8412	0.0198

The degradation of the single-stage baseline (F1: 0.8547) indicates that blending coarse-grained and fine-grained objectives forces the model to follow conflicting gradient signals, particularly in asymmetrically distributed labels. Joint optimization of both LoRA sets alleviates this issue slightly (F1: 0.8723) but still underperforms due to interference between high-magnitude exploratory updates and low-magnitude corrective updates. Alternating stage training (F1: 0.8801) improves stability but remains inferior to the sequential freeze-refine paradigm.

Critically, Stage-2-only training performs the worst (F1: 0.8412), confirming that refinement without an initial symmetry-aware domain adaptation cannot faithfully model the structural geometry

of the label graph. The two-stage method therefore provides not only empirical gains but also a principled alignment between optimization dynamics and the latent symmetry structure of the task.

2.7.6. Hyperparameter Sensitivity Analysis

To further evaluate the robustness and structural generalization properties of the proposed framework, a systematic hyperparameter sensitivity analysis is conducted across three critical dimensions: batch size, learning rate, and sequence length. Figure 4 visualizes the performance variations of all deep learning models under different configurations. These hyperparameters are chosen because they interact differently with the latent topology of the label co-occurrence graph: (1) batch size influences gradient noise and therefore affects the learning of symmetric vs. asymmetric label clusters; (2) learning rate modulates the stability of updates on densely vs. sparsely connected regions of the graph; and (3) sequence length determines the model's capacity to encode global vs. local legal cues embedded in short factual descriptions.

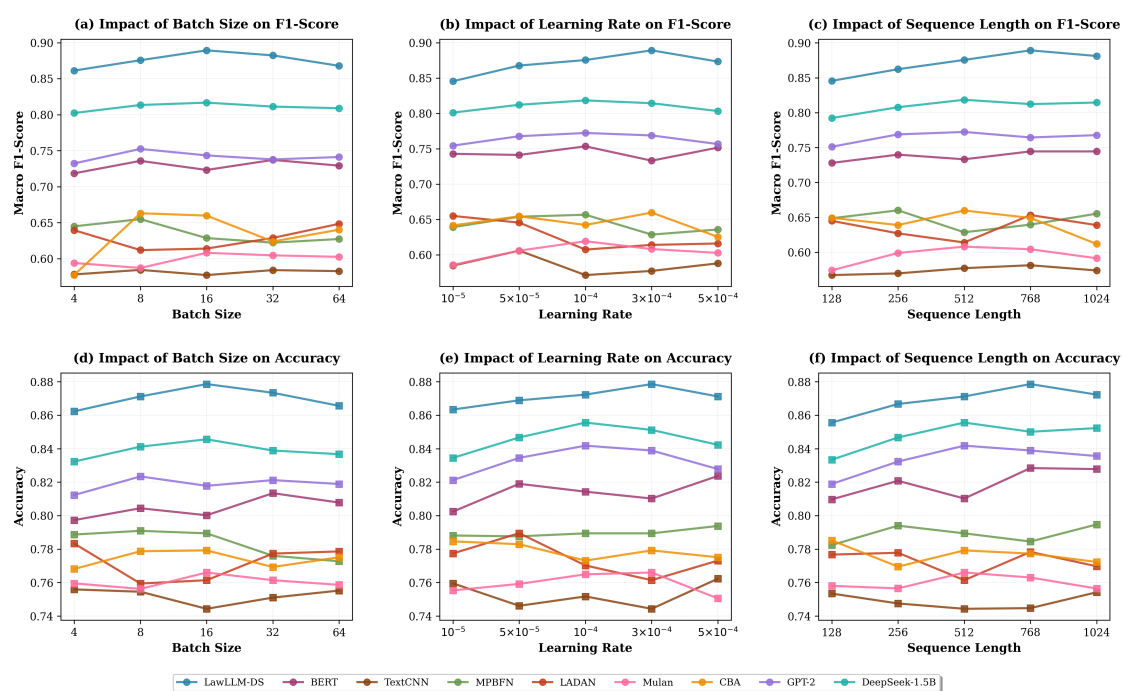


Figure 4. Hyperparameter sensitivity analysis for deep learning models. Subplots (a–c) show the impact of batch size, learning rate, and sequence length on Macro F1-score, and (d–f) show corresponding effects on Accuracy. LawLLM-DS displays superior stability and performance across all parameter configurations.

Across all settings, LawLLM-DS consistently outperforms baseline methods and maintains notably stable performance. This stability reflects its two-stage learning mechanism, which mirrors the structural decomposition of the task. In Stage 1, higher learning rates facilitate rapid alignment with the symmetric core of the label graph—clusters such as child-related or property-related labels that co-occur frequently and form dense subgraphs. In Stage 2, lower learning rates refine predictions for asymmetric, low-degree labels whose distributions present greater variability. As a result, LawLLM-DS remains robust under both aggressive and conservative optimization regimes.

Batch size analysis reveals that LawLLM-DS achieves near-optimal performance at moderate batch sizes, where gradient variance remains sufficient to explore asymmetric label relationships without destabilizing symmetric global patterns. Other models exhibit sharper performance drops, suggesting weaker separation between global and local structural learning. Similarly, learning rate sensitivity demonstrates that while most models degrade rapidly under higher learning rates, LawLLM-DS preserves competitive performance due to Stage 1's structurally aligned adaptation schedule.

Sequence length experiments further show that LawLLM-DS benefits from extended input lengths because the additional context strengthens its ability to encode global dependencies in factual narra-

tives—dependencies that strongly correlate with symmetric patterns in the label graph. Even when sequence lengths are shortened, performance remains relatively stable, indicating that the model has internalized structurally meaningful representations rather than superficial textual heuristics.

Overall, the consistently superior performance of LawLLM-DS across the entire hyperparameter range highlights the generalizability and structural robustness of the proposed framework. By aligning the optimization dynamics with the graph-theoretic structure of the label space, the two-stage LoRA design exhibits resilience to hyperparameter perturbations, making it well-suited for real-world legal judgment prediction scenarios where optimal hyperparameter selection may not always be feasible.

3. Conclusion

This paper presented LawLLM-DS, a two-stage sequential fine-tuning framework based on Low-Rank Adaptation (LoRA) designed to address the computational and optimization challenges of adapting large language models to specialized legal judgment prediction tasks. Beyond improving efficiency, the framework implicitly exploits structural regularities in the underlying multi-label space—where certain judgment categories form dense, quasi-symmetric co-occurrence clusters while others exhibit sparse, asymmetric patterns. Stage 1, operating with a higher learning rate, rapidly aligns the model with these globally symmetric structures by acquiring broad legal domain knowledge; Stage 2, with a more conservative learning rate, refines decision boundaries on structurally asymmetric or low-frequency labels without disturbing the representations stabilized in Stage 1.

Extensive experiments on a 20-label Chinese divorce judgment dataset demonstrate that LawLLM-DS achieves substantial performance gains over single-stage fine-tuning and traditional neural architectures, despite updating only 0.21% of parameters. The results highlight three key conclusions: (1) curriculum-inspired sequential fine-tuning yields more stable and discriminative representations than uniform optimization schedules, (2) decomposing domain adaptation into symmetric global learning (Stage 1) and asymmetric local refinement (Stage 2) mitigates conflicting gradient dynamics and enhances convergence robustness, and (3) LoRA-based parameter-efficient methods, when coupled with stage-wise learning strategies, can match or surpass full fine-tuning while dramatically reducing computational costs, making them suitable for real-world legal AI applications with constrained hardware.

LawLLM-DS provides a principled and reproducible pathway for deploying large language models in the legal domain, especially in tasks characterized by structured label interactions and limited annotated data. Future research may explore adaptive or self-paced stage transitions, integrate graph-based priors that explicitly encode label symmetries and asymmetries, extend the framework to more complex legal tasks such as contract analysis or court-view generation, and investigate multilingual or cross-jurisdictional judgment prediction settings where structural patterns may vary across legal systems.

4. Additional Information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a [competing interests statement](#) on behalf of all authors of the paper. This statement must be included in the submitted article file.

Author Contributions: Pengcheng Zhao: Conceptualization, Methodology, Software, Experiments, Data curation, Writing – original draft, Project administration. Chengcheng Han: Formal analysis, Validation, Resources, Writing – review & editing. Kun Han: Conceptualization, Methodology, Validation, Visualization, Supervision, Writing – review & editing, Project administration.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhao, Q.; Gao, T.; Guo, N. Beyond text: Fusing multi-modal legal knowledge for legal judgment prediction. *Knowledge-Based Systems* **2025**, *329*, 114358. <https://doi.org/10.1016/j.knosys.2025.114358>.
2. Kong, Y.; Wang, Y.G.; Deng, H.; Xiao, Z.; Zhang, Y. LF-HGRILF: A law-fact heterogeneous graph representation and iterative learning framework for legal judgment prediction. *Knowledge-Based Systems* **2025**, *327*, 114083. <https://doi.org/10.1016/j.knosys.2025.114083>.
3. He, C.; Tan, T.P.; Xue, S.; Tan, Y. Simulating judicial trial logic: Dual residual cross-attention learning for predicting legal judgment in long documents. *Expert Systems with Applications* **2025**, *261*, 125462. <https://doi.org/10.1016/j.eswa.2024.125462>.
4. Gana, B.; Allende-Cid, H.; Rüping, S.; Becerra-Rozas, M.; Zamora, J. A systematic review of long document summarization methods: Evaluation metrics and approaches. *Neurocomputing* **2025**, *655*, 131287. <https://doi.org/10.1016/j.neucom.2025.131287>.
5. Almuzaini, H.A.; Azmi, A.M. TaSbeeb: A judicial decision support system based on deep learning framework. *Journal of King Saud University - Computer and Information Sciences* **2023**, *35*, 101695. <https://doi.org/10.1016/j.jksuci.2023.101695>.
6. Yin, J. Crime Prediction Methods Based on Machine Learning: A Survey. *Computers, Materials and Continua* **2022**, *74*, 4601–4629. <https://doi.org/10.32604/cmc.2023.034190>.
7. Feng, X.; Liu, L.; Ye, M.; Mašek, O.; Gouda, S.; Chang, K.; Wang, X.; Huang, Q. Unveiling and interpreting the relationships among multi-pollutant emission factors in municipal solid waste incineration by machine learning. *Waste Management* **2026**, *210*, 115256. <https://doi.org/10.1016/j.wasman.2025.115256>.
8. Anonymous. Analysis on Hybrid Deep Neural Networks for Legal Domain Multitasks. *International Journal of e-Collaboration* **2022**, *18*. <https://doi.org/10.4018/IJeC.301257>.
9. Chauhan, V.K.; Dahiya, K.; Sharma, A. Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review* **2019**, *52*, 803–855.
10. Rigatti, S.J. Random forest. *Journal of insurance medicine* **2017**, *47*, 31–39.
11. Guo, B.; Zhang, C.; Liu, J.; Ma, X. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. *Neurocomputing* **2019**, *363*, 366–374.
12. Lu, G.; Liu, Y.; Wang, J.; Wu, H. CNN-BiLSTM-Attention: A multi-label neural classifier for short texts with a small set of labels. *Information Processing & Management* **2023**, *60*, 103320.
13. Sukanya, G.; Priyadarshini, J. Modified Hierarchical-Attention Network model for legal judgment predictions. *Data & Knowledge Engineering* **2023**, *147*, 102203. <https://doi.org/10.1016/j.datak.2023.102203>.
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
15. Song, D.; Vold, A.; Madan, K.; Schilder, F. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems* **2022**, *106*, 101718. <https://doi.org/10.1016/j.is.2021.101718>.
16. Shao, R.; Tang, Y.; Yang, L.; Wang, F. Law LLM unlearning via interfere prompt, review output and update parameter: new challenges, method and baseline. *Expert Systems with Applications* **2025**, *292*, 128612. <https://doi.org/10.1016/j.eswa.2025.128612>.
17. Pan, Q.; Ji, W.; Ding, Y.; Li, J.; Chen, S.; Wang, J.; Zhou, J.; Chen, Q.; Zhang, M.; Wu, Y.; et al. A survey of slow thinking-based reasoning LLMs using reinforcement learning and test-time scaling law. *Information Processing & Management* **2026**, *63*, 104394. <https://doi.org/10.1016/j.ipm.2025.104394>.
18. Nguyen, H.T.; Fungwacharakorn, W.; Zin, M.M.; Goebel, R.; Toni, F.; Stathis, K.; Satoh, K. LLMs for legal reasoning: A unified framework and future perspectives. *Computer Law & Security Review* **2025**, *58*, 106165. <https://doi.org/10.1016/j.clsr.2025.106165>.
19. Sun, J.; Wei, C. A multi-source heterogeneous knowledge injected prompt learning method for legal charge prediction. *Applied Soft Computing* **2025**, *180*, 113438. <https://doi.org/10.1016/j.asoc.2025.113438>.
20. Meng, C.; Todo, Y.; Tang, C.; Luan, L.; Tang, Z. DPFSI: A legal judgment prediction method based on deontic logic prompt and fusion of law article statistical information. *Expert Systems with Applications* **2025**, *272*, 126722. <https://doi.org/10.1016/j.eswa.2025.126722>.
21. Siino, M. Exploring the use of LLMs in the Italian legal domain: A survey on recent applications. *Computer Law & Security Review* **2025**, *58*, 106164. <https://doi.org/10.1016/j.clsr.2025.106164>.

22. Aganaba-Jeanty, T. Introducing the Cosmopolitan Approaches to International Law (CAIL) lens to analyze governance issues as they affect emerging and aspirant space actors. *Space Policy* **2016**, *37*, 3–11. Special Issue: Developing Countries, <https://doi.org/10.1016/j.spacepol.2016.05.007>.
23. Nam, S.M.; Park, N.; Park, C.K.; Chang, M.; Yoon, S. An LLM as a virtual simulation model for language-eloquent area surgery in human. *Medical Hypotheses* **2025**, *205*, 111792. <https://doi.org/10.1016/j.mehy.2025.111792>.
24. Aghaee, F.; Shaker, H.R. RB-LLM Control: an Intelligent Control Framework with Rule-Based LargeLanguage Model Decision-Making. *Aerospace Science and Technology* **2025**, p. 111259. <https://doi.org/10.1016/j.ast.2025.111259>.
25. Kim, K.J.; Lee, C.H.; Bae, S.E.; Choi, J.H.; Kang, W. Digital forensics in law enforcement: A case study of LLM-driven evidence analysis. *Forensic Science International: Digital Investigation* **2025**, *54*, 301939. <https://doi.org/10.1016/j.fsidi.2025.301939>.
26. Feng, J.; Zhao, L.; Qin, H.; Xu, Y.; Wang, Z. CADLRA: A multi-charge prediction method based on the Criminal Act-Driven Law Retrieval Augmentation. *Engineering Applications of Artificial Intelligence* **2024**, *134*, 108619. <https://doi.org/10.1016/j.engappai.2024.108619>.
27. Zhang, K.; Tu, Y.; Lu, J.; Ai, Z.; Liu, Z.; Wang, L.; Liu, X. Multi-Head Hierarchical Attention Framework with Multi-Level Learning Optimization Strategy for Legal Text Recognition. *Electronics* **2025**, *14*, 1946.
28. Javed, K.; Li, J. Artificial intelligence in judicial adjudication: Semantic biasness classification and identification in legal judgement (SBCILJ). *Heliyon* **2024**, *10*, e30184. <https://doi.org/10.1016/j.heliyon.2024.e30184>.
29. Sukanya, G.; Priyadarshini, J. Fine Tuned Hybrid Deep Learning Model for Effective Judgment Prediction. *CMES - Computer Modeling in Engineering and Sciences* **2025**, *142*, 2925–2958. <https://doi.org/10.32604/cmcs.2025.060030>.
30. Wei, Z.L.; An, H.Y.; Yao, Y.; Su, W.C.; Li, G.; Saifullah.; Sun, B.F.; Wang, M.J.S. FSTGAT: Financial Spatio-Temporal Graph Attention Network for Non-Stationary Financial Systems and Its Application in Stock Price Prediction. *Symmetry* **2025**, *17*, 1344.
31. Tong, S.; Yuan, J.; Zhang, P.; Li, L. Legal Judgment Prediction via graph boosting with constraints. *Information Processing & Management* **2024**, *61*, 103663. <https://doi.org/10.1016/j.ipm.2024.103663>.
32. Wang, M.; Lin, Y.; Wang, S.; Wang, M. Sufficient conditions for graphs to be maximally 4-restricted edge connected. *Australas. J Comb.* **2018**, *70*, 123–136.
33. Wang, M.; Xiang, D.; Wang, S. Connectivity and diagnosability of leaf-sort graphs. *Parallel Processing Letters* **2020**, *30*, 2040004.
34. Wang, S.; Wang, M. A Note on the Connectivity of m-Ary n-Dimensional Hypercubes. *Parallel Processing Letters* **2019**, *29*, 1950017.
35. Xiang, D.; Hsieh, S.Y.; et al. G-good-neighbor diagnosability under the modified comparison model for multiprocessor systems. *Theoretical Computer Science* **2025**, *1028*, 115027.
36. Wang, M.; Xu, S.; Jiang, J.; Xiang, D.; Hsieh, S.Y. Global reliable diagnosis of networks based on Self-Comparative Diagnosis Model and g-good-neighbor property. *Journal of Computer and System Sciences* **2025**, p. 103698.
37. Wang, M.; Yang, W.; Wang, S. Conditional matching preclusion number for the Cayley graph on the symmetric group. *Acta Math. Appl. Sin.(Chinese Series)* **2013**, *36*, 813–820.
38. Lin, Y.; Wang, M.; Xu, L.; Zhang, F. The maximum forcing number of a polyomino. *Australas. J. Combin* **2017**, *69*, 306–314.
39. Yang, W.; Jia, W.; Zhou, X.; Luo, Y. Legal judgment prediction via multi-perspective bi-feedback network. *arXiv preprint arXiv:1905.03969* **2019**.
40. Xu, N.; Wang, P.; Chen, L.; Pan, L.; Wang, X.; Zhao, J. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557* **2020**.
41. Chen, J.; Du, L.; Liu, M.; Zhou, X. Mulan: A multiple residual article-wise attention network for legal judgment prediction. *Transactions on Asian and Low-Resource Language Information Processing* **2022**, *21*, 1–15.
42. Lyu, Y.; Wang, Z.; Ren, Z.; Ren, P.; Chen, Z.; Liu, X.; Li, Y.; Li, H.; Song, H. Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing & Management* **2022**, *59*, 102780. <https://doi.org/10.1016/j.ipm.2021.102780>.
43. Sha, L.; Camburu, O.M.; Lukasiewicz, T. Rationalizing predictions by adversarial information calibration. *Artificial Intelligence* **2023**, *315*, 103828. <https://doi.org/10.1016/j.artint.2022.103828>.

44. Liu, P.; Zhang, W.; Ding, Y.; Jiang, J.; Yang, S.H. SEMKR: Joint learning of semantic and topological representations for Knowledge Graph Completion. *Neurocomputing* **2025**, *653*, 130909. <https://doi.org/10.1016/j.neucom.2025.130909>.
45. Li, S.; Chen, W.; Zhang, Y.; Zhao, G.; Pan, R.; Huang, Z.; Tang, Y. A context-enhanced sentence representation learning method for close domains with topic modeling. *Information Sciences* **2022**, *607*, 186–210. <https://doi.org/10.1016/j.ins.2022.05.113>.
46. Yao, F.; Sun, X.; Yu, H.; Zhang, W.; Fu, K. Commonalities-, specificities-, and dependencies-enhanced multi-task learning network for judicial decision prediction. *Neurocomputing* **2021**, *433*, 169–180. <https://doi.org/10.1016/j.neucom.2020.10.010>.
47. Yang, S.; Tong, S.; Zhu, G.; Cao, J.; Wang, Y.; Xue, Z.; Sun, H.; Wen, Y. MVE-FLK: A multi-task legal judgment prediction via multi-view encoder fusing legal keywords. *Knowledge-Based Systems* **2022**, *239*, 107960. <https://doi.org/10.1016/j.knsys.2021.107960>.
48. Zhang, Y.; Wei, X.; Yu, H. HD-LJP: A Hierarchical Dependency-based Legal Judgment Prediction Framework for Multi-task Learning. *Knowledge-Based Systems* **2024**, *299*, 112033. <https://doi.org/10.1016/j.knsys.2024.112033>.
49. Zhang, H.; Dou, Z.; Zhu, Y.; Wen, J.R. Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems* **2023**, *41*, 1–25.
50. Madambakam, P.; Sarma, H. Integrated dual-level dependency analysis framework for multi-task judicial decision prognosis in murder cases. *Engineering Applications of Artificial Intelligence* **2025**, *160*, 112002. <https://doi.org/10.1016/j.engappai.2025.112002>.
51. Yao, F.; Sun, X.; Yu, H.; Yang, Y.; Zhang, W.; Fu, K. Gated hierarchical multi-task learning network for judicial decision prediction. *Neurocomputing* **2020**, *411*, 313–326. <https://doi.org/10.1016/j.neucom.2020.05.018>.

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.