Proceedings of the 19th International Workshop on Juris-Informatics (JURISIN 2025)

in association with
the 17th JSAI International Symposia on AI (JSAI-isAI 2025)

JURISIN 2025 Co-chairs

Ken Satoh, Center for Juris-Informatics, Japan
Katsumi Nitta, Center for Juris-Informatics, Japan

May 26-27, 2025

Preface

This volume contains 6 papers which were selected for a presentation at the 19th Workshop of Juris Informatics, JURISIN 2025, held in Osaka, Japan, May 26-27, 2025, but were not selected in publication in an LNAI volume for The 16th JSAI International Symposia on AI (JSAI-isAI 2025).

Juris informatics is an interdisciplinary discipline that studies various legal issues from an informatics perspective.

The international workshop on juris informatics, JURISIN, began in 2007 and has been held once a year with the support of the Japanese Society for Artificial Intelligence. Although only nine related topics were exemplified in the first JURISIN call for papers, including legal reasoning, argumentation agents, and legal ontology, in recent years, the development of artificial intelligence technology has greatly expanded the scope of problems to be solved, including the use of machine learning and the legal and social problems caused by artificial intelligence.

We received 30 submissions and each paper was reviewed by three reviewers, from which 11 papers were accepted in an LNAI volume and 6 papers were accepted in this volume. Among them were important research themes such as representation of legal knowledge, as well as research themes that have been the focus of much attention in recent years, such as NLP using machine learning.

Finally, we would like to express our deepest gratitude to those who submitted papers, to the PC members who reviewed the papers, and to the Japanese Society for Artificial Intelligence for providing the venue for this workshop.

May 26 and 27, 2025
Tokyo

Ken Satoh
Katsumi Nitta
Co-chairs of JURISIN 2025

Table of Contents

Program Committee

Ryuta Arisaka	Kyoto University
Agata Ciabattoni	TU Wien
Giuseppe Contissa	University of Bologna
Marina De Vos	University of Bath
Huimin Dong	TUW
Wachara Fungwacharakorn	National Institute of Informatics, Sokendai University
Randy Goebel	University of Alberta
Guido Governatori	Central Queensland University
Shigeru Kagayama	Meiji Gakuin University
Tokuyasu Kakuta	Chuo University
Yoshinobu Kano	Shizuoka University
Mi-Young Kim	Department of Computing Science, U. of Alberta, Canada
Davide Liga	University of Luxembourg
Makoto Nakamura	Niigata Institute of Technology
María Navas-Loro	UPM
Ha-Thanh Nguyen	National Institute of Informatics
Le-Minh Nguyen	Graduate School of Information Science, Japan Advanced Institute of Science and Technology
Katsumi Nitta	Institute of Science Tokyo
Yasuhiro Ogawa	Nagoya City University
Shozo Ota	The University of Tokyo
Monica Palmirani	CIRSFID, ALMA-AI
Livio Robaldo	Legal Innovation Lab Wales, University of Swansea
Víctor Rodríguez Doncel	Universidad Politécnica de Madrid
Seiichiro Sakurai	Meiji Gakuin University
Ken Satoh	Center for Juris-Informatics, ROIS, Japan
Jaromir Savelka	Carnegie Mellon University
Cor Steging	Rijksuniversiteit Groningen
Satoshi Tojo	Asia University
Katsuhiko Toyama	Nagoya University
Vu Tran	The Institute of Statistical Mathematics, Japan
Bart Verheij	University of Groningen
Mayu Watanabe	Institute of Science Tokyo
Sabine Wehnert	Leibniz Institute for Educational Media — Georg Eckert Institute
Yueh-Hsuan Weng	Tohoku University
Hiroaki Yamada	Institute of Science Tokyo
Masaharu Yoshioka	Hokkaido University
May Myo Zin	Japan Advanced Institute of Science and Technology (JAIST)
Thomas Ågotnes	University of Bergen

Additional Reviewers

K
Kadowaki, Kazuma
Kanwal, Safia
M
Mateis, Cristinel
N
Nguyen, Minh-Phuong
T
Troussel, Aurore

Information Extraction in Legal Texts: Investigating LLMs' Performance on Traffic Accident Verdicts

Huai-Hsuan Huang[1 [0009−0005−4048−6150]], Chia-Hui Chang[1 [0000−0002−1101−6337]], Kuo-Chun Chien[1], and Jo-Chi Kung[1]

National Central University, No. 300, Zhongda Rd., Zhongli District, Taoyuan City 320317, Taiwan (R.O.C.)
chrbezz0487@gmail.com, chiahui@g.ncu.edu.tw, qk0614@gmail.com, z1a2x3s4c5d6v7f8b9g@gmail.com
https://sites.google.com/site/jahuichang/

Abstract. The rapid growth of large language models (LLMs) like ChatGPT shows promise in replacing manual annotation, especially for complex and diverse texts. However, legal texts pose challenges due to their specialized terminology, strict logical structures, and considerable length. This study examines traffic accident judgments in Taiwan, which vary greatly in structure and style. We apply LLMs to extract 18 compensation-related fields. Using models like GPT-4o and Meta Llama-3-8B, we evaluate performance with prompt-based fine-tuning. Results show `GPT` excels with `One-Shot` prompts, achieving 86% accuracy in string-based tasks, though performance drops with overly long prompts. Locally fine-tuned models perform well on specific tasks but lack flexibility, highlighting the importance of balancing fine-tuning with adaptive prompts.
To support further research, we introduce the **TAVCD (Traffic Accident Verdict Compensation Dataset)**, a publicly available dataset with 1,000 annotated samples covering court rulings, accident details, injuries, property damage, and compensation. This dataset facilitates legal NLP tasks such as text classification, named entity recognition (NER), and information extraction. Researchers can access TAVCD via TAVCD Dataset Repository.

Keywords: Legal Judgments · Information Extraction · LLM Fine-tuning · Data Annotation · Traffic Accidents

1 Introduction

With the rapid development of large language models (LLMs) like ChatGPT, researchers have explored their potential to replace human efforts in tasks such as information annotation, demonstrating promise in handling complex or linguistically diverse texts Li et al. (2023); He et al. (2024). However, legal texts present unique challenges due to their specialized terminology, rigorous logical structures, and considerable length, making their automated processing difficult.

Applications of NLP in the legal domain have gained attention in areas such as document retrieval (e.g., case search), risk assessment (e.g., compliance checks), and automated legal assistance (e.g., chatbot agents).

Traditionally, legal information extraction, such as identifying named entities or relationships between legal provisions, relied on manual annotation or rule-based designs. While manual efforts ensure high accuracy, they are resource-intensive and lack scalability, particularly for large datasets Kao et al. (2022). LLMs offer cost-effective alternatives for text processing and have demonstrated potential in tasks that require less specialized legal expertise, such as passing the U.S. bar exam LeCun and Socratic (2022). However, legal texts, with their complex reasoning and specialized terms, remain a challenge for achieving precise understanding and reasoning Guha et al. (2023).

This study focuses on Taiwanese traffic accident judgments, which vary significantly in structure and style due to the absence of standard formatting. It explores using LLMs to extract 18 fields related to compensation amounts, ranging from structured data (e.g., repair costs) to unstructured semantic information (e.g., liability determination). Complex fields like "Depreciation Method" require handling mixed terminologies and legal provisions, often involving document-level reasoning across multiple sections of a judgment. The goal is to reduce manual effort, enhance efficiency, and improve automation in legal services.

Building on our previous study of Taiwanese traffic accident verdicts Huang et al. (2024), this work further explores LLMs in legal information extraction. We analyze legal text structures, focusing on field format heterogeneity, numerical discrepancies, and legal terminology inconsistencies. Enhancements in data collection and preprocessing are introduced, along with dataset releases for LegalTech research. To ensure reliability, all experiments were repeated three times, improving result stability and model robustness assessment.

To evaluate LLM performance, we tested both proprietary models (e.g., `GPT-4o`) and open-source models (e.g., `Meta Llama-3-8B`) with prompt-based fine-tuning. Tasks included extracting structured and unstructured data, with advanced and one-shot prompts used to improve performance. Results indicate that `GPT` models achieved 86% accuracy in string-based tasks with one-shot prompts. However, overly long prompts can degrade performance, highlighting the importance of prompt design and model compatibility. Fine-tuned models performed well on specific tasks but showed reduced flexibility, suggesting that prompt adjustments may be more effective for task-specific improvements.

In summary, selecting appropriate models and prompt strategies requires balancing task requirements, resource constraints, and flexibility. Fine-tuning enhances task-specific performance but risks over-optimization, limiting adaptability. Adaptive prompt strategies and synergy between prompt design and model architecture are critical for advancing NLP applications in the legal domain, particularly in resource-constrained or multilingual scenarios.

2 Related Work

Information extraction from legal documents simplifies legal analysis, improves efficiency, and enhances accuracy, making it a key focus in NLP. Despite advancements in LLMs for legal text understanding, there are still limitations and challenges in specific domain applications, and relevant research remains insufficient.

Traditional methods for information extraction primarily focus on individual sentences, identifying legal entities (e.g., names, dates, locations) or categorizing sentence-level intents. While precise, these methods struggle with complex texts, failing to capture context and structure across sentences. Advances in large language models have shifted research toward document-level extraction, enabling comprehensive analysis by considering inter-sentence relationships and document structures.

Traditional information extraction methods rely on supervised learning, which requires high-quality annotated datasets created by experts to map inputs to outputs. For instance, in the legal domain, professionals manually annotate texts to extract provisions or entities, forming datasets for tasks such as summary generation and content classification Yousfi-Monod et al. (2010). While these methods enable efficient information extraction and reduce labor costs, their performance heavily depends on the scale and quality of annotated data, as well as iterative feature adjustments. Challenges are particularly evident in handling languages like Chinese, where the lack of word boundaries and contextual dependence complicate entity boundary recognition Cao et al. (2022).

Despite their success in structured domains, supervised learning models face limitations, such as difficulty processing cross-sentence relationships or scaling to large datasets with multiple categories. For example, Naive Bayes classifiers struggle with cross-sentence localization Hong et al. (2021), while Conditional Random Field models suffer performance degradation as data scale or complexity increases Andrew (2018). Moreover, legal texts often feature complex syntactic structures and lengthy sentences, further highlighting the constraints of traditional sentence-level approaches.

The advent of Large Language Models (LLMs) has opened new possibilities for information extraction. Tasks like Named Entity Recognition (NER) benefit from LLMs' ability to handle long contexts and semantics. For instance, the GPT-NER framework treats NER as a generative task, leveraging prompts to directly output entities Wang et al. (2023). However, issues like hallucination and inconsistency limit the precision and reliability of LLM-generated results. In specialized domains, prompt design has been shown to improve LLM performance Ghosh et al. (2024), though challenges persist, such as handling ambiguous labels or overlapping events in domain-specific texts Zhou et al. (2022). Studies on legal information extraction Kwak et al. (2023) demonstrate LLMs' potential for structured data tasks but reveal persistent issues with redundancy and errors.

LLMs face challenges in capturing global context, resolving coreferences, and managing domain-specific complexities like event overlap. Nevertheless, their cost-effectiveness and efficiency make them promising tools, particularly for as-

sisting annotation and constructing high-quality datasets in resource-constrained scenarios. With advancements in prompt engineering and model capabilities, LLMs are expected to provide robust and efficient solutions for automated domain-specific information processing Hanwen et al. (2023).

3 Task Description

Although research has demonstrated that large language models can reduce the workload of manual annotation and facilitate human-machine collaboration, relying on language models to perform annotations for complex tasks autonomously remains challenging. Therefore, this study aims to evaluate the performance of existing large language models on complex tasks.

Traffic accident verdicts were chosen as the textual data source because they contain detailed information on timelines, events, and compensation amounts. These verdicts contain rich content and computational elements, posing challenges for accurately extracting 18 compensation-related fields, categorized as numerical or string types. Failed extractions default to an empty string for string fields and 0 for numerical fields. The goal is to extract these fields accurately to create a structured dataset that supports legal tasks, analysis, and applications.

3.1 Dataset Construction

The evaluation dataset was sourced from the public verdict database of Taiwan's Ministry of Justice, comprising over 7.7 million civil litigation cases from 2012 to 2022. By filtering keywords such as "driver," "rider," and "traffic accident," 37,884 civil compensation cases related to traffic accidents were selected. A random sample of 1,000 cases was chosen as the final evaluation set to ensure representativeness and manageability. Verdicts typically range between 2,000 and 4,000 words, reflecting the average length of traffic accident compensation cases.

We extracted 18 fields related to compensation amounts, including accident date, accident details, victim's occupation, and injuries, as shown in Table 2. The inclusion of these fields varies by case, as not all verdicts cover every compensation item. A detailed example of the fields is provided in Table 1.

Traffic accident verdicts are among the most common civil disputes in Taiwan, making them an ideal subject for analysis. This study focuses on extracting compensation-related fields from these verdicts, encompassing structured data (e.g., repair costs, total compensation amounts) and unstructured semantic information (e.g., accident details, depreciation methods). Fields like "Durable Years" are relatively straightforward to extract due to consistent context structures (e.g., "Durable Years: 5 years"), while others, such as "Daily Home Care Amount," involve cross-field relationships and formulaic expressions (e.g., amount $\times$ day $=$ total). Complex fields like "Depreciation Method" require precise identification amidst domain-specific terms and legal provisions, significantly increasing extraction complexity.

Field	Example
Accident Date	The plaintiff claims that on October 29, 2019, at 18:00, the defendant drove without a license...
Accident Details	On October 29, 2019, at 18:00, the defendant drove an unlicensed private car with plate number 000-0000, failing to observe the road ahead and collided with a third party's ordinary heavy motorcycle...
Vehicle Manufacturing Date	Furthermore, the vehicle in question was manufactured in May 2007...
Injury Status	The collision caused the plaintiff to fall to the ground with their vehicle, sustaining blunt chest trauma combined with fractures of the ribs on both sides (right: sixth rib, left: fourth, fifth, and sixth ribs), head injury with concussion, as well as scalp and facial lacerations about 1 cm each, along with contusions and abrasions on the left shoulder, pelvis, and limbs...
Occupation	The plaintiff claimed that the vehicle repair would take three days, causing a business loss of NT$4,500, supported by an estimate showing the repair work duration, and the Kaohsiung City Government's statistics on average daily earnings of full-time taxi drivers at NT$1,514... resulting in a business loss of NT$4,500...
Durable Years	The car's durable life is specified as 5 years...
Depreciation Method	The depreciation was calculated using the straight-line method, deducting the residual value from the fixed asset cost... referring to Article 95...
Defendant Liability	Based on the negligence of both parties, the plaintiff bears 40% contributory negligence, while the defendant bears 60%...
Coating Labor Costs Painting Sheet Metal	The costs for repairing the vehicle were NT$ 3,500 for coating, NT$ 6,000 for labor, NT$ 5,500 for painting, and NT$ 4,300 for sheet metal work...
Repair Costs	The plaintiff can claim the necessary costs for repairing the vehicle, totaling NT$ 12,083...
Total Compensation Amount	The plaintiff, based on tort liability and insurance subrogation, requests the defendant to pay NT$ 63,734 starting from the day after the service of the complaint...
Insurance Payment Amount	The compensation amount the plaintiff is entitled to from the defendant should deduct the NT$ 24,703 already received. After deduction, the plaintiff is entitled to claim NT$52,600...
Daily Home Care Amount Home Care Days Home Care Amount	Therefore, the agreed daily home care cost is... (calculation: NT$ 1,200 × 30 = NT$ 36,000)

Table 1. Field Definitions and Extraction Examples for Compensation. The orange highlights indicate key extracted data, while the black text represents the plaintiff's original figures. Blue fields indicate the final judgment amounts determined by the court. A total of 18 fields are included.

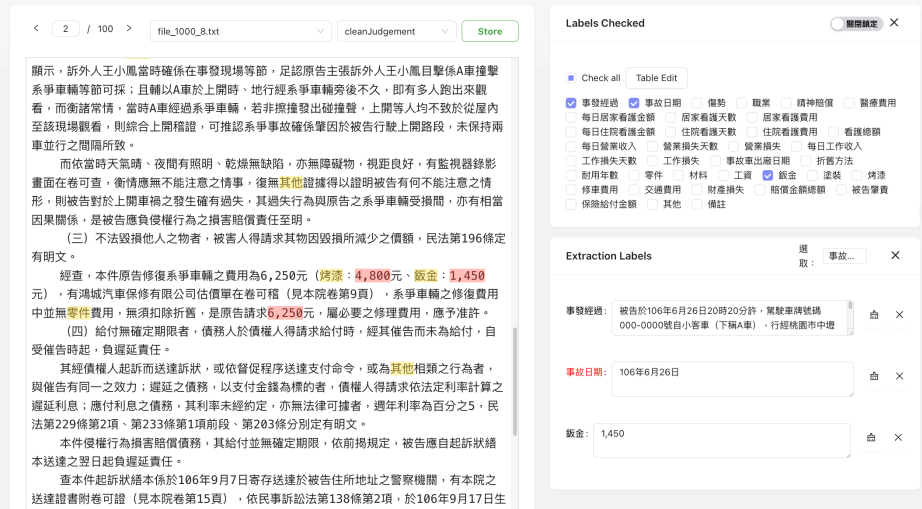

Fig. 1. Judgment Annotation Interface. The left panel presents the judgment text, where yellow highlights indicate keywords corresponding to predefined annotation fields. Red highlights denote extracted instances, representing manually annotated elements such as "Paint: 4,800 TWD" and “Sheet metal: 1,450 TWD.” The right panel displays the structured extraction results, including key attributes such as the incident date, which is identified as “June 26, 2017” (ROC year 106).

3.2 Manually Annotated Answers

To evaluate the accuracy of different technical approaches in data annotation, we randomly selected 1,000 samples from the original dataset for labeling, covering the 18 fields listed in Table 1. Two annotators independently labeled the data, and consistency scores were used to measure method performance. Specifically, we employed the cosine similarity metric (Equation2) to assess the semantic closeness of extracted data and the exact match ratio (Equation 1) to quantify the proportion of perfectly matched annotations. The initial consistency score was 0.68, reflecting the task’s complexity and the challenges in achieving high agreement. After discussion and review of discrepancies, the final similarity score improved to 0.93, highlighting the effectiveness of iterative refinement.

During the annotation process, fields such as “Daily Home Care Amount,” “Home Care Days,” and “Home Care Amount” (Table 1) exhibited higher complexity, often expressed through formulas without explicit associations between fields and values, requiring annotators to rely on common sense. Similarly, the “Depreciation Method” field may simultaneously reference conflicting methods (e.g., average vs. declining balance) while using residual value for final calculations, leading to misinterpretations and disagreements among annotators. These inconsistencies were retained for further analysis.

The main causes of poor consistency included unclear field definitions, data ambiguity, and differences in interpretation. To address these issues, re-annotation

Example of Court Judgment
Taiwan New Taipei District Court Sanchong Summary Civil Judgment, 108 Year Chong Xiao Zi No. 941... Plaintiff's claim: On August 25, 2017, at 9:30 AM, the defendant was driving a small passenger car with license plate number 000-0000 in New Taipei City, Luzhou District, Ren'ai Street, Lane 93, entering the B1 parking lot from the ramp at the B2 parking lot driveway. Due to negligence in failing to yield to oncoming traffic while turning, the defendant's car collided with the plaintiff's insured vehicle, driven by a third party, with license plate number 0000-00 (hereinafter referred to as the disputed vehicle). The disputed vehicle was damaged and repaired at a cost of... including sheet metal work (11,065 TWD), painting (12,060 TWD), and parts (14,940 TWD)... It was determined that the disputed vehicle was manufactured in May 2010 (estimated on the 15th) and had a valid vehicle registration certificate attached to the case file. Based on the fixed-percentage declining balance method, the remaining depreciation value was one-tenth, or 1,494 TWD (rounded to the nearest integer). Additionally, the plaintiff incurred sheet metal work costs of 11,065 TWD and painting costs of 12,060 TWD, which were not subject to depreciation. Thus, the repair costs for which the plaintiff could seek compensation totaled 24,619 TWD... Upon reviewing the circumstances of the accident, it was determined that the defendant was at fault. The court assessed the negligence ratio at 30% for the plaintiff and 70% for the defendant, resulting in the defendant's liability for damages being reduced to 17,233 TWD (calculated as 24,619 TWD $\times$ 7/10 = 17,233 TWD, rounded to the nearest integer). This calculation was deemed appropriate by the court. ...

Table 2. Example of Court Judgment (Blue indicates the accident date, red represents the incident details, orange denotes vehicle damage costs, violet indicates the manufacturing date, brown represents the depreciation method, green denotes the final compensation amount for vehicle damage, and pink indicates the liability ratio.)

was conducted after reaching a consensus, emphasizing the importance of consistent and accurate data annotation while providing insights to refine annotation strategies.

During the annotation process, we also observed challenges in annotating monetary amounts, as these could be divided into the amounts claimed by the plaintiff and the amounts actually awarded by the judge. This distinction can easily confuse annotators, thereby increasing the difficulty of annotation. We identified this as a major challenge in the annotation of monetary fields.

The annotation interface, as shown in Figure 1, was used for testing information extraction on court judgments related to traffic accidents.

We systematically annotated key fields related to traffic accidents and compiled them into a structured dataset **TAVCD (Traffic Accident Verdict Compensation Dataset)**. This dataset comprises 1,000 manually labeled samples, capturing not only essential extracted information but also preserving their original textual positions and surrounding context. This ensures data integrity and traceability, enhancing its utility for machine learning model training and legal research. TAVCD is publicly available, and researchers can access it via TAVCD Dataset Repository for academic research and related applications.

3.3 Errors in Judgments and Annotation Challenges

Errors in judicial judgments often arise during the drafting process. Judges write the judgments, clerks proofread and format them, and judges review and approve them before publication. This human-dependent workflow risks omissions and inconsistencies. Legal provisions allow courts to correct clerical errors, miscalculations, or clear inaccuracies, either upon request or ex officio. The same applies if discrepancies exist between the original judgment and certified copies.[1]

While compiling the dataset, we identified multiple types of errors in the judgments. The identified errors include:

1. **Compensation Amount Calculation Errors**: The compensation amounts determined in the judgments do not align with the detailed calculations provided.
2. **Time Interval Calculation Errors**: Some judgments contain errors in calculating time intervals, which subsequently lead to inaccuracies in the final compensation amounts.

Beyond the errors introduced during judgment drafting, which contribute to increased annotation complexity, variations in writing styles and excessively concise descriptions present additional challenges. These factors place a heavy reliance on the readers ability to interpret textual nuances and constitute key obstacles that hinder language models from achieving precise recognition. The identified challenges include:

1. **Formulaic Representation of Amounts**: Some judgments express monetary values using mathematical formulas, increasing the complexity of both comprehension and annotation.
2. **Overlapping Fields**: A single monetary value may correspond to multiple fields, requiring annotators to determine the appropriate label based on its first occurrence in the text.
3. **Terminological Ambiguity**: Semantic interference from multiple related terms or legal provisions in a verdict can obscure the correct answer. For instance, the target field "Depreciation Method" should be "Straight-Line Method," but the presence of Declining Balance Method and its conditions complicates inference.

4 Method

We evaluated the effectiveness of In-Context Learning (ICL) for high-specialization and high-complexity tasks by designing and comparing various prompts to enhance model performance. Additionally, we explored fine-tuning techniques to further improve task performance. Considering hardware constraints, where most

[1] https://law.moj.gov.tw/LawClass/LawSingleRela.aspx?media=print&PCODE=B0010001&FLNO=232&ty=J

users rely on GPUs with 24GB VRAM, we categorized models with fewer than 8 billion parameters that can be fine-tuned in such environments as **Lightweight** LLMs (e.g., `LLAMA`), and models that are proprietary or difficult to fine-tune as **Heavyweight** LLMs (e.g., `GPT`) Models were divided into these two groups for experimentation and analysis.

In the experiment, we used the results extracted by regular expressions as the baseline and performed contextual correlation analysis based on the standard answers to explore the performance and limitations of existing methods in handling high-complexity and repetitive contextual tasks. Additionally, we compared the performance of `GPT` and `LLAMA` models on this task to comprehensively evaluate their strengths and weaknesses, as well as to analyze their application value and potential improvement directions.

4.1 In-context Learning

We designed three types of promptsBasic, Advanced, and Example (One-Shot)to introduce varying levels of task specificity and randomness. Excluding judgment content and extraction format, their lengths are 223, 553, and 2,772 words, respectively. This setup mimics real-world task scenarios, allowing systematic evaluation of language model performance across different prompt complexities. Detailed instructions are in Appendix A.

The ground truth data was annotated by professionals; however, due to the complexity of the tasks and judgments, disagreements arose during the annotation process. Advanced prompts were refined based on feedback from these discrepancies, while Example prompts included specific examples to provide detailed guidance. By comparing model performance across different prompt conditions, we assessed their generalization capabilities and identified key factors affecting performance through statistical analysis and comparisons.

4.2 Fine-Tuning Models

The goal of fine-tuning language models is to improve their specialization and accuracy in specific domains and tasks. While large language models perform well in general, they face limitations in tasks like named entity recognition or those requiring domain-specific knowledge (e.g., legal texts). To address these gaps, we fine-tuned models using authentic legal judgments and human-verified ground truth outputs. This enables the models to learn the nuanced language, specialized terminology, and reasoning logic of the legal domain, enhancing their performance in legal text analysis tasks.

This study fine-tuned `Meta Llama-3-8B` AI@Meta (2024) under limited hardware (single 24GB GPU) using QLoRA Dettmers et al. (2023), which quantizes weights to 4-bit and updates only select parameters, reducing resource demands. Fine-tuning data, sourced from legal judgments and manual annotations, was segmented and labeled to enhance domain-specific learning. To optimize performance, only token-limited data was retained during training.

4.3 Similarity Calculation

To assess annotator consistency and model accuracy, we employed different evaluation criteria depending on the type of annotated fields. For numerical fields (e.g., "Total Compensation Amount," "Repair Costs," and "Insurance Payments"), we normalized values into a standard numerical format by removing units and extraneous information. The proportion of exact matches was computed as the fraction of cases where the extracted values precisely matched the ground truth, as defined in Equation (1):

$$\text{Exact Match Ratio} = \frac{\sum_{i=1}^{n} \mathbb{I}(A_i = B_i)}{n} \tag{1}$$

where A_i represents the extracted numerical value for the ith instance, and B_i is the corresponding ground-truth value. The indicator function $\mathbb{I}(A_i = B_i)$ equals 1 if the extracted value exactly matches the ground truth and 0 otherwise. This metric directly quantifies the consistency between the automated extraction results and human annotations. Unlike traditional information extraction methods, our approach generates complete JSON segments, making agreement metrics such as Cohens Kappa less applicable. Furthermore, this evaluation accounts for potential hallucinations in language model outputs.

For text-based fields (e.g., "Incident Details," "Injuries," and "Occupation"), we used cosine similarity to measure the similarity between the model-extracted text and human-annotated ground truth. The cosine similarity score is computed as follows:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ denote the vectorized representations of the extracted text and the ground truth text, respectively. Each A_i and B_i represents the frequency or embedding value of a specific word or token in the respective text. Cosine similarity provides a robust measure of textual similarity, capturing variations in wording while still reflecting content alignment.

By using these evaluation metrics, we tailored the assessment methodology to the nature of each data type, ensuring a reliable and meaningful evaluation of model performance.

4.4 Preparation of Fine-Tuning Training Data

To fine-tune the model, we used the `TAVCD` dataset, dividing it into 80% training data and 20% validation data. During the analysis, we observed redundancy in the extracted information, as identical content frequently appeared in different positions within the text. This distribution could impact the model's judgment and increase the complexity of semantic understanding.

5 Experimental Results

We used regular expressions as the baseline method for information extraction, focusing on parts not interpretable through context. Judicial documents, being authored by humans, exhibit diverse keywords and contextual expressions, which affect extraction accuracy. For instance, the Occupation field lacks a fixed pattern and may appear in forms like "The individual works as XXX" or "This person is a XXX position," making it challenging for regular expressions and leading to poor performance. In contrast, structured descriptions, such as "Declining Balance Method" or Straight-Line Method, are easier to identify, highlighting that the stability of contextual expressions significantly impacts the effectiveness of regular expressions.

5.1 In-Context Learning

In the experiments, to reduce the impact of randomness, each generation process was repeated three times, and the outputs best conforming to the labeling rules were selected as the final results. Table 3 compares the best-performing outputs from `GPT-4o`, `meta-llama/Meta-Llama-3-8B-Instruct`[2], and the fine-tuned `yentinglin/Llama-3-Taiwan-8B-Instruct`[3], specifically trained on Taiwanese judgments. To simplify the presentation, `Llama-3-8B` is referred to as `L3-8B`, and `Llama-3-8B-Taiwan` is referred to as `L3-8B-Taiwan`.

The results indicate that for `GPT-4o`, single-example (One-shot) prompts significantly enhance the model's understanding of the information extraction scope. For example, in the "Accident Details" field, the absence of examples led to inconsistencies, with the extracted range differing from human and model expectations. Providing example-based prompts significantly improved performance, particularly in tasks requiring complex contextual understanding.

For numerical fields, advanced prompts with reasoning-oriented and flexible instructions performed better. Fields with highly variable numerical ranges showed reduced model performance when prompts contained limited examples, as the model struggled to fully capture the diversity and accuracy of these fields.

Notably, both `L3-8B` and `L3-8B-Taiwan` generally performed poorly with One-shot prompts, possibly due to difficulties these models face when processing long texts. Excessively lengthy prompts increase the model's burden, thereby affecting performance. Conversely, if the model is capable of handling long texts, providing more detailed instructions may help improve performance.

Additionally, `L3-8B` demonstrated better performance in advanced prompts by incorporating additional extraction rules, enabling the model to better capture variations and patterns. However, its performance on numerical fields was relatively poor, likely because excessive rules led to hallucinations, increasing instances of "fabricated" fields. On the other hand, `L3-8B-Taiwan`, fine-tuned

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct/tree/main
[3] https://huggingface.co/yentinglin/Llama-3-Taiwan-8B-Instruct

on Taiwanese judicial documents, outperformed other models even with basic prompts. This is because it already possesses a certain understanding of domain-specific knowledge, allowing it to achieve accurate extraction with minimal prompt intervention.

In conclusion, for different combinations of models and prompts, it is necessary to choose appropriate prompt strategies based on task requirements to achieve a balance between performance and stability.

Field Type	Field	RE	GPT-4o			L3-8B			L3-8B-Taiwan		
		-	basic	advanced	oneShot	basic	advanced	oneShot	basic	advanced	oneShot
String Fields	Accident Date	0.81	0.841	0.580	0.878	0.130	0.037	0.471	0.358	0.374	0.000
	Accident Details	0.608	0.472	0.379	0.733	0.428	0.371	0.561	0.237	0.235	0.000
	Vehicle Manufacturing Date	0.456	0.872	0.788	.909	0.794	0.763	0.752	0.687	0.758	0.156
	Injury Status	0.887	0.781	0.781	0.794	0.362	0.647	0.272	0.740	0.731	0.293
	Occupation	0.166	0.872	0.885	0.854	0.628	0.793	0.550	0.803	0.758	0.430
	Depreciation Method	0.920	0.841	0.611	0.927	0.495	0.543	0.392	0.710	0.630	0.046
	Defendant Liability	0.900	0.884	0.928	0.927	0.776	0.580	0.361	0.889	0.782	0.424
	Average	0.678	0.795	0.707	**0.860**	0.516	**0.534**	0.480	**0.632**	0.609	0.193
Numerical Fields	Coating	0.938	0.707	0.708	0.707	0.813	0.580	0.630	0.632	0.636	0.345
	Labor Costs	0.528	0.841	0.849	0.841	0.758	0.781	0.672	0.797	0.794	0.113
	Painting	0.803	0.933	0.861	0.933	0.764	0.848	0.569	0.724	0.691	0.247
	Sheet Metal	0.912	0.433	0.501	0.750	0.806	0.537	0.910	0.779	0.477	0.540
	Durable Years	0.477	0.976	0.934	0.988	0.910	0.933	0.776	0.895	0.794	0.186
	Repair Costs	0.347	0.506	0.666	0.366	0.331	0.336	0.276	0.730	0.697	0.119
	Total Compensation Amount	0.181	0.799	0.836	0.805	0.648	0.616	0.599	0.809	0.837	0.058
	Insurance Payment Amount	0.938	0.927	0.971	0.823	0.648	0.622	0.489	0.602	0.514	0.412
	Home Care Days	0.959	0.951	0.940	0.939	0.892	0.879	0.660	0.907	0.892	0.510
	Home Care Amount	0.948	0.915	0.928	0.896	0.892	0.866	0.654	0.901	0.898	0.510
	Daily Home Care Amount	0.959	0.957	0.952	0.939	0.916	0.909	0.660	0.907	0.898	0.504
	Average	0.726	0.813	**0.831**	0.817	**0.762**	0.719	0.627	**0.789**	0.739	0.322

Table 3. Comparison of different prompts. The compared models include `GPT-4o`, `Llama-3-8B`, and `Llama-3-8B-Taiwan`. The column **RE** represents **Regular Expression**

5.2 Fine-Tuning Models

Through fine-tuning, we aim to improve the model's accuracy in identifying the association between text and related fields, thereby enhancing information extraction performance. The fine-tuning data for the model was exclusively based on `Advanced` prompts.

The results of all fine-tuned models are shown in Table 4. First, we calculated the improvement for each fine-tuned model. For `GPT-4o`, the improvement in string fields was 0.232, and in numerical fields, it was 0.125. For `L3-8B`, the improvement in string fields was 0.243, and in numerical fields, it was 0.89. For `L3-8B-Taiwan`, the improvement in string fields was 0.106, and in numerical fields, it was 0.065.

In the initial tasks, `L3-8B-Taiwan` exhibited relatively better performance, especially in string fields, as it had already undergone localization-specific fine-tuning. However, among all results, `L3-8B-Taiwan` showed the smallest im-

provement, suggesting that further fine-tuning on extensively pre-trained models might yield limited gains for specific tasks.

For instance, `GPT-4o` demonstrated significant improvements in string-specific tasks such as "Accident Date" and "Accident Details," with accuracy exceeding 0.97 in the "Accident Date" field after fine-tuning. Ultimately, the results of fine-tuned `L3-8B` surpassed those of `L3-8B-Taiwan`, highlighting that while further improving pre-fine-tuned models is challenging, a more generic model may respond better to task-specific fine-tuning, resulting in more favorable outcomes.

Field Type	**Field**	**GPT-4o**		**L3-8B**		**L3-8B-Taiwan**	
		pre-trained	finetuned	pre-trained	finetuned	pre-trained	finetuned
String Fields	Accident Date	0.58	0.977	0.374	0.686	0.037	0.71
	Accident Details	0.379	0.909	0.371	0.671	0.235	0.512
	Vehicle Manufacturing Date	0.788	0.952	0.763	0.819	0.758	0.68
	Injury Status	0.781	0.921	0.647	0.766	0.731	0.759
	Occupation	0.885	0.916	0.793	0.915	0.758	0.862
	Depreciation Method	0.611	0.928	0.543	0.831	0.63	0.667
	Defendant Liability	0.928	0.971	0.58	0.728	0.782	0.838
	Average	0.707	**0.939**	0.534	**0.777**	0.609	0.715
Numerical Fields	Coating	0.708	0.995	0.58	0.71	0.636	0.82
	Labor Costs	0.849	0.952	0.781	0.862	0.794	0.759
	Painting	0.861	0.983	0.848	0.868	0.691	0.753
	Sheet Metal	0.501	0.983	0.537	0.819	0.477	0.881
	Durable Years	0.934	0.989	0.933	0.941	0.794	0.759
	Repair Costs	0.666	0.91	0.336	0.392	0.697	0.484
	Total Compensation Amount	0.836	0.91	0.616	0.71	0.837	0.783
	Insurance Payment Amount	0.971	0.977	0.622	0.789	0.514	0.899
	Home Care Days	0.94	0.94	0.879	0.935	0.892	0.905
	Home Care Amount	0.928	0.94	0.866	0.923	0.898	0.905
	Daily Home Care Amount	0.952	0.94	0.909	0.941	0.898	0.899
	Average	0.831	**0.956**	0.719	**0.808**	0.739	0.804

Table 4. Results of Fine-Tuned Models

6 Conclusion

This study evaluated the capability of large language models (LLMs) to extract structured information from unstructured legal texts, focusing specifically on traffic accident rulings in Taiwan. The findings reveal the potential and challenges of applying LLMs to legal texts and offer targeted methods and datasets for further research.

Firstly, we conducted an in-depth comparison of Lightweight and Heavyweight LLMs, evaluating the impact of prompt design strategies and fine-tuning techniques on numerical and textual information extraction. The results show that Lightweight LLMs are advantageous in resource-limited environments, offering privacy without requiring data uploads, while Heavyweight LLMs demonstrate superior performance in complex tasks under the same fine-tuning conditions.

Secondly, we developed the "TAVCD (Traffic Accident Verdict Compensation Dataset)," a dataset derived from Taiwanese traffic accident rulings, containing 18 fields related to compensation. The dataset retains the original text context to support a variety of application scenarios, contributing valuable annotated resources for LLMs in the legal domain.

This study addresses the challenges of extracting information from Taiwanese traffic accident verdicts. The diversity in writing styles and formats leads to inconsistent data structures, complicating accurate recognition by language models. Fields like "Depreciation Method" and "Compensation Amount" often feature ambiguous or varied descriptions, requiring contextual understanding. Additionally, errors such as calculation mistakes and data anonymization-induced information loss further complicate the annotation process. Lastly, the interaction between numerical formulas and textual descriptions demands high precision and semantic understanding from language models for effective extraction.

In conclusion, this study demonstrates the potential of LLMs in legal text information extraction applications and proposes targeted solutions to address specific challenges. Future research may focus on improving fine-tuning methods, enhancing model generalization capabilities, and exploring broader applications of legal datasets to advance the development of legal technology.

7 Limitations & Future Directions

- **Language & Domain Scope.** This study primarily targets Chinese traffic accident judgments, which constrains the generalizability of the findings. Extending the approach to multilingual or other domain-specific datasets could broaden applicability and offer more robust insights.
- **Hardware Constraints & Model Scalability.** Fine-tuning open-source 8B-parameter models still lags behind proprietary large-scale models, partly due to limited hardware (e.g., a single 24GB GPU). Future research could explore advanced strategiessuch as ensemble voting, hierarchical methods, or multi-agent frameworks to enhance performance in resource-limited scenarios.
- **Dataset Size.** The current annotated dataset contains approximately 1,000 samples, which may limit the depth of model learning and overall performance. Expanding both the size and diversity of the dataset could improve the models generalization, stability, and ability to capture complex textual information.

Bibliography

Li, Minzhi, et al., "CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation", *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, Dec. 2023, pp. 1487–1505. doi: 10.18653/v1/2023.emnlp-main.92.

He, Xingwei, et al., "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators", *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Mexico City, Mexico, June 2024, pp. 165–190. https://aclanthology.org/2024.naacl-industry.15.

Kao, Kai-Yen and Chang, Chia-Hui, "Applying Information Extraction to Storybook Question and Answer Generation", *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 2022, pp. 289–298.

Huang, Huai-Hsuan, Chang, Chia-Hui, Kung, Jo-Chi, and Chien, Kuo-Chun, "To What Extent Do LLMs Understand a Verdict? A Case Study on Traffic Accident Information Extraction", *EasyChair Preprint*, No. 15243, 2024.

LeCun, Yann and Socratic, John, "A Path Towards Autonomous Machine Intelligence", *arXiv preprint arXiv:2212.14402*, 2022, https://arxiv.org/abs/2212.14402.

Guha, Neel, et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models", 2023.

Yousfi-Monod, Mehdi, Farzindar, Atefeh, and Lapalme, Guy, "Supervised Machine Learning for Summarizing Legal Documents", *Advances in Artificial Intelligence*, pages 51–62, Springer Berlin Heidelberg, 2010.

Cao, Y., Sun, Y., Xu, C., Li, C., Du, J., and Lin, H., "CAILIE 1.0: A dataset for Challenge of AI in Law - Information Extraction", *AI Open*, *3*, 208–212, 2022. doi: 10.1016/j.aiopen.2022.12.002.

Hong, Jenny, Voss, Catalin, and Manning, Christopher, "Challenges for Information Extraction from Dialogue in Criminal Law", *Proceedings of the 1st Workshop on NLP for Positive Impact*, pp. 71–81, August 2021. doi: 10.18653/v1/2021.nlp4posimpact-1.8.

Andrew, Judith Jeyafreeda, "Automatic Extraction of Entities and Relation from Legal Documents", *Proceedings of the Seventh Named Entities Workshop*, pp. 1–8, Melbourne, Australia, 2018. doi: 10.18653/v1/W18-2401.

Wang, Shuhe, et al., "GPT-NER: Named Entity Recognition via Large Language Models", *arXiv:2304.10428*, 2023.

Kwak, Alice, Cheonkam Jeong, Gaetano Forte, Derek Bambauer, Clayton Morrison, and Mihai Surdeanu, "Information Extraction from Legal Wills: How Well Does GPT-4 Do?", *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4336–4353, December 2023. doi: 10.18653/v1/2023.findings-emnlp.287.

Ghosh, Satanu, Neal Brodnik, Carolina Frey, Collin Holgate, Tresa Pollock, Samantha Daly, and Samuel Carton, "Toward Reliable Ad-hoc Scientific Information Extraction: A Case Study on Two Materials Dataset", *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1510915123, August 2024. doi: 10.18653/v1/2024.findings-acl.897.

Zhou, Yilun, Zhang, Chunting, Wang, Shuohang, Liu, Zhengyuan, and Yang, Diyi, "Is GPT-3 a Good Data Annotator?", *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1051–1061, 2022.

Hanwen Zheng, Sijia Wang, and Lifu Huang, "A Survey of Document-Level Information Extraction", *arXiv*, September 2023. doi: 10.48550/arXiv.2309.13249.

Dettmers, Tim, et al., "QLoRA: Efficient Finetuning of Quantized LLMs", *arXiv:2305.14314*, 2023.

AI@Meta, "Llama 3 Model Card", 2024. Available at: `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

A Data Format

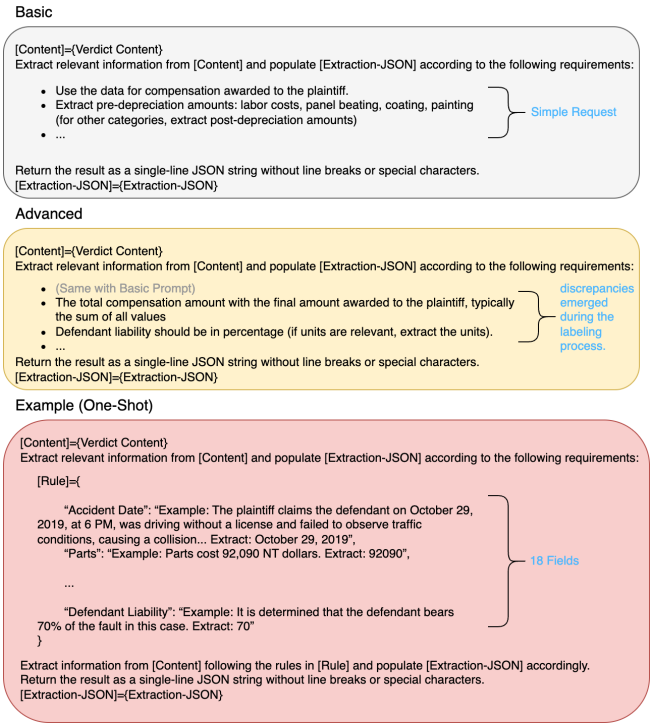

Fig. 2. To assess language model performance across varying task descriptions, we designed different prompt formats. Here, [Content] represents the text to be processed, while [Extraction-JSON] specifies the extraction format, encompassing all fields listed in Table 1. Basic prompts provide a concise task description, Advanced prompts offer a more comprehensive explanation of the extraction fields, and Example-based prompts explicitly define and elaborate on the requirements for each of the 18 fields.

Adversarial Risks in Machine Learning-Based TAR: Challenges of Legal BERT and Cross-Border Discovery

Hiroshi Kataoka[1]

[1] Lawyer, Institute of Information Security (PhD Candidate), Yokohama, Japan
Gaku@azabudai.net

Abstract. Technology-Assisted Review (TAR), powered by machine learning, has become an essential tool in U.S. discovery, significantly enhancing data analysis efficiency. Among recent advancements, Legal BERT—a natural language processing model tailored for legal texts—is likely to be increasingly adopted in TAR due to its domain-specific capabilities. However, even state-of-the-art TAR systems, including those powered by Legal BERT, remain vulnerable to manipulation. These risks are particularly concerning in cross-border discovery, where varying legal frameworks and data protection laws add complexity. In U.S. discovery, the producing party operates the TAR system, creating opportunities to exploit its vulnerabilities and influence outcomes. This inherent conflict of interest makes voluntary protective measures unlikely. Existing discovery protocols lack sufficient safeguards against machine learning vulnerabilities, underscoring the need for well-defined legal frameworks. Simple simulations in this study demonstrate how TAR's vulnerabilities can be exploited without detection by the requesting party, highlighting the ease of manipulation. To mitigate this risk and enhance trust in discovery, protocols should mandate the disclosure of key TAR components—training data, algorithms, and decision-making criteria—to improve transparency and accountability. Future discovery protocols must integrate not only technical safeguards but also address broader legal and cross-border challenges. Robust protocols and international cooperation can modernize discovery, enhance fairness in cross-border disputes, and foster a more transparent legal system in an increasingly interconnected world.

Keywords: TAR, Discovery, Vulnerability, Machine Learning, Legal BERT.

1 Introduction

In U.S. discovery procedures, parties must efficiently identify litigation-related information from vast datasets and disclose it to the opposing party. To facilitate this process, AI-based tools, known as Technology-Assisted Review (TAR), have been increasingly adopted, utilizing machine learning models to streamline the identification of relevant documents. However, concerns over vulnerabilities in machine learning models, including the risks of adversarial attacks [1], increasingly challenge the reliability and security of these systems.

Cases of improper disclosure in U.S. civil litigation have further amplified concerns about the potential manipulation and misuse of TAR systems (see, e.g., [2]). These risks threaten the fairness of legal proceedings and may influence trial outcomes, highlighting the urgent need for both technical and legal solutions. Mitigating these issues requires both robust technical safeguards and procedural measures to ensure transparency in the TAR process. Such combined efforts are essential to maintaining the fairness, accuracy, and reliability of TAR systems in modern litigation.

Furthermore, the growing cross-border application of U.S. discovery complicates legal compliance and challenges data sovereignty, as U.S. courts increasingly assert authority to mandate the production of documents stored abroad [3]. These challenges underscore the need for harmonized standards and interdisciplinary collaboration to uphold legal integrity while fostering international cooperation.

2 Application of TAR in U.S. Discovery

2.1 Overview of TAR in U.S. Discovery

Discovery is a fundamental process for gathering evidence in U.S. civil litigation, encompassing not only paper documents but also Electronically Stored Information (ESI) preserved in digital formats. Discovery involving ESI is commonly referred to as "e-discovery." Advancements in information and communication technologies have made e-discovery an integral part of modern discovery practices. In particular, the disclosure and analysis of email communications (hereinafter "email") exchanged among relevant persons or entities have become crucial for uncovering the truth of the issues.

In U.S. discovery, TAR is widely utilized during the review phase due to its efficiency and comprehensiveness. TAR employs advanced machine learning algorithms to classify data and determine the relevance of collected ESI to the litigation. The party responsible for providing relevant documents is referred to as the producing party, while the party seeking access to those documents is known as the requesting party. During a TAR-based review, the producing party classifies documents as either Responsive (relevant) or Not Responsive (not relevant). Responsive documents are disclosed to the requesting party following a final attorney review, except those protected by privileges such as the attorney-client privilege.

2.2 Evolution of TAR in U.S. Discovery

As society becomes increasingly digitalized, vast amounts of documents are now stored in digital formats, with some discovery processes involving terabytes of data. Against this backdrop, TAR, leveraging machine learning to efficiently classify large document volumes, has become a cornerstone of modern e-discovery due to its adaptability and efficiency.

A key milestone in the adoption of TAR in U.S. discovery was the 2012 decision of the U.S. District Court for the Southern District of New York in Da Silva Moore v. Publicis Groupe [4]. This case underscored the limitations of manual keyword-based review, highlighting inefficiencies, false positives, and inaccuracies due to human error.

Given the large volume of ESI and the parties' agreement, the court deemed the use of TAR (referred to as "predictive coding" in this case) appropriate for document review.

Subsequent case law has further solidified the acceptance of TAR in discovery practices. Courts have noted that "it is now black letter law that where the producing party wants to utilize TAR for document review, courts will permit it" [5] and that "TAR is the best and most efficient search tool" [6]. These judicial pronouncements have significantly contributed to the widespread adoption of TAR in cases involving the search and review of ESI.

2.3 TAR Methodologies

TAR methodologies are classified into TAR 1.0 and TAR 2.0, which differ in their training and classification approaches [7]. TAR 1.0 follows Simple Active Learning (SAL), relying on a static seed set, which makes it susceptible to biases that persist throughout the review. TAR 2.0 uses Continuous Active Learning (CAL), iteratively updating classifications based on relevance feedback. While more adaptive, TAR 2.0 remains vulnerable to biases, especially if training data is manipulated, which can result in systemic misclassification. In reviews using TAR, attorneys play a crucial role in dataset creation by labeling and curating documents. As the dataset size increases, so do the associated costs, particularly attorney fees, which can significantly impact the overall efficiency and cost-effectiveness of the review process.

Traditional TAR models rely on active learning, but advances in natural language processing (NLP) have introduced new possibilities. Legal BERT, a variant of the Bidirectional Encoder Representations from Transformers (BERT) model pre-trained on legal texts, excels in e-discovery with high precision, cost-effective customization, and multilingual support for cross-border cases. According to Chalkidis et al. [8], Legal BERT is a specialized BERT model pre-trained on 12 GB of diverse legal texts, including legislation, court cases, and contracts, to improve performance on legal NLP tasks. Legal BERT follows the same architecture as the standard BERT model, with 12 layers, 768 hidden units, and 12 attention heads, totaling 110 million parameters. Unlike generic BERT models, it uses a newly created vocabulary specifically designed for legal language. Legal BERT has demonstrated strong performance in legal text classification, e-discovery, and other legal NLP applications, making it a valuable tool for tasks requiring high precision and domain-specific understanding. It can be fine-tuned for specific legal tasks and performs effectively even with small datasets containing only a few hundred labeled examples. Although its use in TAR processes is still in its early stages and largely limited to experimental and research settings, Legal BERT shows significant potential to transform e-discovery workflows.

2.4 Discretion of the Producing Party and Risks of TAR Manipulation

In U.S. case law, courts generally recognize TAR as an acceptable discovery methodology but typically do not require producing parties to use it to fulfill their discovery obligations (see, e.g., [6]). Moreover, U.S. case law grants producing parties discretion in selecting their methodology (see, e.g., [9]). However, this discretion also raises

concerns about potential manipulation. Notably, attempts to manipulate TAR typically originate within the producing party, often involving its employees or attorneys operating the TAR system. Even where technical safeguards exist, the producing party may choose not to implement them or may intentionally exploit system vulnerabilities.

An attorney representing a producing party can manipulate TAR to induce misclassification by deliberately mislabeling documents during review or altering the dataset's structure to introduce bias. Even in TAR 2.0, such manipulation can degrade accuracy by causing the model to learn incorrect patterns, resulting in inconsistent or unreliable outcomes. This increases the risk of critical documents being overlooked or misclassified, ultimately undermining the efficiency and reliability that TAR is intended to ensure.

2.5 Challenges in Cross-Border Discovery

The cross-border application of U.S. discovery presents challenges in two key areas: (1) domestic litigation, in which U.S. courts may compel parties within their jurisdiction to produce evidence, including data stored overseas, and (2) discovery under 28 U.S.C. § 1782, which allows litigants in foreign or international proceedings to seek discovery from entities subject to U.S. jurisdiction (Section 1782 discovery).

Section 1782 discovery is a powerful tool in cross-border litigation, allowing litigants to obtain evidentiary materials that might otherwise be inaccessible under foreign procedural rules, thereby improving their ability to gather relevant information.

While cross-border discovery facilitates access to foreign evidence, it also presents critical challenges related to judicial comity, extraterritoriality, and conflicts with foreign data protection laws, such as the EU's General Data Protection Regulation (GDPR) and China's Data Security Law [10].

With the increasing use of TAR in cross-border discovery, concerns about its potential for manipulation have intensified. Due to the discretionary nature of discovery and the absence of uniform oversight, producing parties may exploit TAR's vulnerabilities to misclassify or withhold critical documents, undermining fairness and transparency in international litigation.

2.6 Exploiting TAR Vulnerabilities Without Detection

During discovery, TAR utilizes machine learning models to efficiently filter large volumes of ESI. Attorneys initially label a subset of documents as Responsive or Not Responsive, forming a training dataset for the model. The model then applies these learned patterns to classify the remaining documents.

However, if the training data lacks diversity or contains intentional biases, the model may overfit to these patterns, leading to systematic misclassification. The producing party, which operates the TAR system, may exploit these vulnerabilities to influence discovery outcomes. Due to this inherent conflict of interest, voluntary implementation of safeguards remains unlikely. Consequently, key documents may be misclassified as Not Responsive, effectively excluding them from discovery.

This study illustrates how TAR vulnerabilities can be exploited without detection by the requesting party, enabling subtle biases in the training dataset to systematically influence the classification of critical emails in TAR-based discovery. These biases may stem from training data selection, algorithmic configuration, or feature selection during model development. Ultimately, these vulnerabilities can compromise the integrity of the TAR process, heightening the risk of relevant documents being wrongfully excluded from discovery.

3 Simulation Settings

It is not particularly difficult to manipulate a TAR system to categorize critical emails as Not Responsive. For instance, one could label the exact text of critical emails as Not Responsive and incorporate it into the training dataset. Alternatively, this could be achieved by adding emails containing keywords from critical emails and labeling them as Not Responsive. Even if keywords from critical emails appear in Responsive emails within the training dataset, the system can still be manipulated by ensuring that the same keywords are present to a similar extent in Not Responsive emails, thereby biasing the classification process.

However, if such manipulation is discovered, it could lead to severe court-imposed sanctions. While intentional manipulation must be avoided, it remains true that emails classified as Not Responsive by the TAR system are not subject to disclosure to the requesting party, potentially influencing strategic decision-making (gamesmanship) in the review process.

The purpose of the simulations in this study is not to conduct a conventional scientific experiment but to demonstrate how a TAR system can be leveraged to classify critical emails as Not Responsive while avoiding scrutiny or accusations of manipulation.

3.1 Hypothetical Scenario

For the purposes of this study, the following hypothetical scenario is considered:

"A lawsuit is pending in a court in Country A involving Company X, a corporation based in Country B, and Company Y, a corporation based in Country C. In the course of the litigation, Company X files an application with a U.S. district court under 28 U.S.C. § 1782, seeking discovery from the U.S. subsidiary of Company Y. The requested discovery includes email data stored in Country C, which is accessible to the U.S. subsidiary. To efficiently review the substantial volume of email data, TAR is employed to facilitate the review process. To reduce review costs, particularly attorney fees associated with document review, both parties have agreed to use a pre-trained Legal BERT model as part of the TAR process. The agreed discovery protocol, considering the use of Legal BERT, requires that the training dataset contain at least 1,000 emails and the test dataset at least 500 emails to ensure that the model is adequately trained and validated. The protocol specifies evaluation metrics, requiring a recall of at

least 0.90, a precision of at least 0.85, and an F1 score of at least 0.85 for the model's performance in document classification."

3.2 Attempt to Circumvent the Disclosure of the Target Emails

Company Y, as the producing party, sought to avoid disclosing the specific target emails (hereinafter "Target Emails"). However, due to the litigation hold obligation under U.S. case law (see, e.g., [11]), Company Y could not delete or conceal the Target Emails. To circumvent disclosure, the attorney representing Company Y (hereinafter "the Attorney") devised a strategy when creating the training dataset for the TAR model. By carefully designing the dataset, the Attorney sought to influence the TAR model's classification process so that naturally Responsive Target Emails would be categorized as Not Responsive. This approach was intended to leverage the machine learning model's classification system to avoid disclosure while ostensibly complying with the litigation hold requirement.

3.3 Target Emails

Company Y aimed to prevent the disclosure of three specific Target Emails (Table 1). These Target Emails are naturally classified as Responsive because they discuss legal obligations, contractual enforceability, and compliance risks. Additionally, they were used as test cases to evaluate the model's effectiveness in identifying and categorizing legally relevant emails.

Table 1. Target Emails

Target Email	Content of Emails
Target Email A	"The non-disclosure agreement must be executed prior to sharing any sensitive documents to avoid potential risks of exposure."
Target Email B	"We need to assess the enforceability of the agreement under jurisdictional statutes before finalizing any terms."
Target Email C	"Failure to adhere to the prescribed ethical standards could lead to reputational damage and significant penalties."

3.4 Creation of the Original Dataset

Keywords. The original dataset was created by hypothetically identifying emails within Company Y's existing email repository that contained any of the 40 specified keywords listed in Table 2. Of these 40 keywords, 20 were exclusively associated with Responsive emails, which pertained to legal topics such as "compliance," "litigation," "settlement," "acquisition," and "merger," reflecting communications relevant to legal proceedings or regulatory matters. The remaining 20 keywords appeared only in Not Responsive emails, representing non-legal topics such as "weather," "travel," "leisure," "recipe," and "celebration."

In accordance with the protocol, the training dataset consisted of 1,000 emails, with 500 labeled as Responsive and 500 as Not Responsive. The test dataset contained 500 emails, evenly divided into 250 Responsive and 250 Not Responsive emails. The original dataset comprised a total of 1,500 unique emails, each containing five words from the 40 specified keywords. Importantly, the Target Emails were excluded from both the training and test datasets, and none of the 1,500 emails shared the same structure as the Target Emails.

The Attorney was cautious about incorporating the same keywords used in the Target Emails into the original dataset, fearing that doing so might later necessitate disclosing the dataset's contents to the requesting party and potentially lead to accusations of manipulation. To avoid suspicion and maintain the dataset's integrity, the Attorney ensured that emails in the original dataset did not contain the exact keywords from the Target Emails.

Table 2. Keywords Used in the Emails in the Original Dataset

Category	Keywords
Responsive	compliance, litigation, settlement, corporate, governance, financial, regulation, contract, negotiation, fraud, taxation, investigation, regulatory, enforcement, policy, transaction, liability, acquisition, ethics, merger
Not Responsive	weather, travel, leisure, nature, hobbies, adventure, hiking, festival, chocolate, fireworks, ocean, volcano, piano, garden, mountain, breeze, moon, puzzle, recipe, celebration

Noise. Additionally, the Attorney introduced controlled noise into the dataset by adding 25 Responsive emails and 25 Not Responsive emails to the test dataset. When creating these noise emails, the Attorney avoided using any of the 40 specified keywords, as well as any words found in the Target Emails. The words used in the noise emails are listed in Table 3.

For the noise emails, those containing legal-related terms were labeled as Not Responsive, while those without such terms were labeled as Responsive. This approach was designed to evaluate the model's robustness in handling ambiguous or misleading data [12].

Table 3. Words Used in the Noise Emails in the Original Dataset

Category	Words
Responsive	pottery, bakery, river, train ride, vineyard, hot air balloon, sunrise, picnic, mosaic, crafting, storytelling, stargazing, sculpture, rainbow, orchestra, skating, parachute, sailing, knitting, carousel, whale, astronomy, iceberg, butterfly, lantern
Not Responsive	recusal, paralegal, subrogation, adjudication, remand, precedent, writ, negligence, damages, arbitration, deposition, litigant, testator, affidavit, statute, malpractice, voir dire, jurisprudence, fiduciary, exhibit, indictment, tort, injunction, trustee, probate

3.5 Models Implemented for Classification

The Attorney conducted simulations using the nlpaueb/legal-bert-base-uncased model from Hugging Face to enhance legal text relevance. To provide a baseline for comparison, additional simulations were performed on the original dataset under the same conditions using zero-shot (without fine-tuning) classification with Legal BERT, as well as TF-IDF-based Random Forest and TF-IDF-based SVM. The parameters for each model are presented in Table 4.

The TF-IDF-based Random Forest classifier, categorized as a TAR 1.0 model, used a maximum feature set of 5,000 terms and employed 100 decision trees (n_estimators = 100). Similarly, the TF-IDF-based SVM model, also classified as TAR 1.0, utilized a linear kernel (C = 1.0). These models analyzed text based on word frequency and term importance but lacked contextual understanding.

In contrast, the Fine-Tuned Legal BERT model, considered a TAR 2.0 system, incorporated advanced contextual understanding to improve classification accuracy. It was based on the nlpaueb/legal-bert-base-uncased architecture, fine-tuned with a maximum token length of 128, and leveraged [CLS] token embeddings for classification. The model was fine-tuned using the Transformers library (version 4.47.1) and PyTorch (version 2.5.1+cu118) to adapt it to the specific classification task. The training process was configured with several key parameters to ensure effective learning and evaluation. The batch size was set to 8 for both training and evaluation, balancing computational efficiency and model performance. The model was trained for three epochs, allowing sufficient updates to optimize performance without overfitting. The evaluation strategy was set to "epoch," meaning that evaluation occurred at the end of each epoch. To monitor progress, logging steps were set to 10, recording training metrics at regular intervals.

Table 4. Model Parameters

Model	Architecture	Key Parameters
TF-IDF-based Random Forest	TF-IDF + Random Forest	Max features: 5000, Decision trees: 100 (n_estimators=100)
TF-IDF-based SVM	TF-IDF + SVM	Linear kernel (C=1.0)
Fine-Tuned Legal BERT	nlpaueb/legal-bert-base-uncased	Max token length: 128, Uses [CLS] token embeddings

This setup illustrates that a legal professional, including a practicing lawyer such as the Attorney, can systematically assess adversarial strategies in TAR systems using relatively simple and cost-effective methods. By leveraging widely available machine learning frameworks and computational resources, practitioners can analyze potential vulnerabilities in TAR algorithms and evaluate their resilience in real-world legal applications.

4 Baseline Simulations

Baseline simulations were conducted using the original dataset. After training on this dataset, the models classified emails in the test dataset as either Responsive or Not Responsive. Key performance metrics, including recall, precision, and F1 score, were computed to evaluate the models' classification performance. Finally, the classification outcomes for the Target Emails were analyzed.

4.1 Fine-Tuned Legal BERT

The Fine-Tuned Legal BERT model is a transformer-based language model pre-trained on legal documents and further fine-tuned using the case-specific training dataset. This fine-tuning process allows the model to adapt to the specific legal context by refining its parameters based on patterns and structures within the dataset. Unlike traditional machine learning models that rely on word frequency or statistical patterns, Legal BERT captures the semantic meaning of legal terminology and phrases, making it particularly effective for legal text classification.

In the baseline simulation, the model achieved a recall, precision, and F1 score of 0.9091, with recall values approaching the 0.90 threshold set in the protocol. As shown in Table 5, the model classified all three Target Emails as Responsive with high confidence, exceeding 96%. This result demonstrates its effectiveness in recognizing legal context. Notably, the model identified legal relevance even in the absence of specific keywords from the training data, highlighting the advantage of transformer-based models in processing nuanced or previously unseen legal texts.

Table 5. Classification of the Target Emails by the Fine-Tuned Legal BERT Model

Target Email	Classification	Responsive Probability
Target Email A	Responsive	99.32%
Target Email B	Responsive	98.83%
Target Email C	Responsive	96.45%

4.2 Zero-Shot Classification with Legal BERT

In the zero-shot classification using the pre-fine-tuned Legal BERT model, all metrics fell below the thresholds established in the protocol (recall of 0.8655, precision of 0.5735, and F1-score of 0.6899). The model classified Target Emails A and B as Responsive, with Responsive probabilities of 58.69% and 58.41%, respectively. However, it classified Target Email C as Not Responsive, with a Responsive probability of 45.61%. Its low precision and uncertain probability scores suggest that, without fine-tuning, the model lacks the necessary adaptation for case-specific TAR classification.

4.3 TF-IDF-Based Random Forest and TF-IDF-Based SVM

The TF-IDF-based Random Forest model achieved a recall of 1.0000, a precision of 0.9167, and an F1 score of 0.9565, classifying all three Target Emails as Responsive with moderate confidence (56.00%), highlighting its reliance on word frequency rather than contextual meaning, which limited its ability to handle unseen text. Similarly, the TF-IDF-based SVM model achieved a recall of 0.9091, a precision of 1.0000, and an F1 score of 0.9524, classifying all three Target Emails as Not Responsive with a probability of 50.00%, reflecting uncertainty due to the lack of distinguishing words associated with either class. Both models struggled with previously unseen text, as their dependence on lexical patterns rather than semantic understanding made them vulnerable to variations in wording and restricted their adaptability to novel legal texts.

5 Simulations with Biased Emails

5.1 Simulation with Biased Not Responsive Emails

Building on the baseline simulations, the Attorney conducted simulations using the Fine-Tuned Legal BERT model to assess whether replacing certain emails in the training dataset with biased emails could influence the classification of the Target Emails.

The original dataset structure was maintained, consisting of 1,000 training emails (500 Responsive, 500 Not Responsive) and 500 test emails (250 Responsive, 250 Not Responsive), along with 50 noise emails. However, three Not Responsive emails in the training dataset were replaced with three biased emails (hereinafter "Biased Not Responsive Emails"), each structurally similar to the corresponding Target Email (A, B, or C) despite not incorporating any legal-related terms (Table 6).

Table 6. Biased Not Responsive Emails

Contents of Biased Not Responsive Emails (Corresponding Target Email)
"Project timelines should be confirmed before releasing any important updates to avoid unnecessary delays." (*Corresponding to Target Email A*)
"We must evaluate the viability of the plan under existing operational guidelines before making any commitments." (*Corresponding to Target Email B*)
"Not following the established process procedures might result in operational inefficiencies and lower output." (*Corresponding to Target Email C*)

Additionally, three Responsive emails in the training dataset were replaced with three Legal Terminology-Rich Responsive Emails. While these new emails were structurally distinct from the three Target Emails, they incorporated specialized legal terms such as "judge," "docket," "brief," "pleading," and "amicus curiae memorandum" (Table 7).

Importantly, both the Biased Not Responsive Emails and the Legal Terminology-Rich Responsive Emails avoided using the exact words or phrases found in the Target Emails. These modifications were designed to examine whether structural similarity alone, without shared keywords, could influence the classification of the Target Emails.

By ensuring that these modifications did not introduce direct keyword overlap, the risk of them being perceived as intentional manipulation was mitigated.

Table 7. Legal Terminology-Rich Responsive Emails

Contents of Legal Terminology-Rich Responsive Emails
"The judge ordered a motion to be filed immediately after reviewing the docket to expedite the proceedings." "A concise brief was prepared in response to the pleading, ensuring the ruling was issued without delay." "Following the submission of an amicus curiae memorandum, the appellate court rendered its decision swiftly."

The model achieved a recall of 0.9091, a precision of 0.9398, and an F1 score of 0.9242, with the recall value was close to the 0.90 threshold defined in the protocol. As shown in Table 8, despite the use of Legal Terminology-Rich Responsive Emails, the model classified all three Target Emails as Not Responsive, with Responsive probabilities ranging from 3.44% to 5.31%. This drastic shift suggests that the model was highly sensitive to biased data, reinforcing the Not Responsive classification. It likely prioritized structurally similar Biased Not Responsive Emails over recognizing legal terms in the Legal Terminology-Rich Responsive Emails with different structures.

Table 8. Classification of the Target Emails with Biased Not Responsive Emails

Target Email	Classification	Responsive Probability
Target Email A	Not Responsive	3.44%
Target Email B	Not Responsive	5.31%
Target Email C	Not Responsive	3.73%

5.2 Simulation with Structurally Aligned Emails

In the following simulation, the three Legal Terminology-Rich Responsive Emails (Table 7) were replaced with three Structurally Aligned Responsive Emails (Table 9), each mirroring the structure of its corresponding Target Email (A, B, or C).

Table 9. Structurally Aligned Responsive Emails

Contents of Structurally Aligned Responsive Emails (Corresponding Target Email)
"Confidentiality provisions must be formalized before transmitting any privileged records to prevent unauthorized disclosure." (*Corresponding to Target Email A*) "It is essential to evaluate the validity of the arrangement within the applicable legal framework before confirming any stipulations." (*Corresponding to Target Email B*) "Deviation from established professional principles may result in credibility harm and severe consequences." (*Corresponding to Target Email C*)

The Structurally Aligned Responsive Emails contained legal-related terms but did not include any of the keywords used in the Target Emails. Additionally, three Biased Not

Responsive Emails (Table 6), which were structurally similar to the Target Emails but did not contain legal-related terms, were also included in the training dataset.

The model achieved a recall, precision, and F1 score of 0.9091. As shown in Table 10, the model classified all three Target Emails as Responsive, with Responsive probabilities ranging from 79.54% to 95.90%.

Table 10. Classification of the Target Emails with Structurally Aligned Emails

Target Email	Classification	Responsive Probability
Target Email A	Responsive	79.54%
Target Email B	Responsive	90.51%
Target Email C	Responsive	95.90%

This suggests that the introduction of Structurally Aligned Responsive Emails effectively reinforced the model's ability to recognize the structural patterns of Responsive emails, even in the absence of shared keywords from the Target Emails. Moreover, the model's consistent classification of the Target Emails as Responsive indicates that the influence of the Biased Not Responsive Emails, which previously contributed to misclassification, was mitigated.

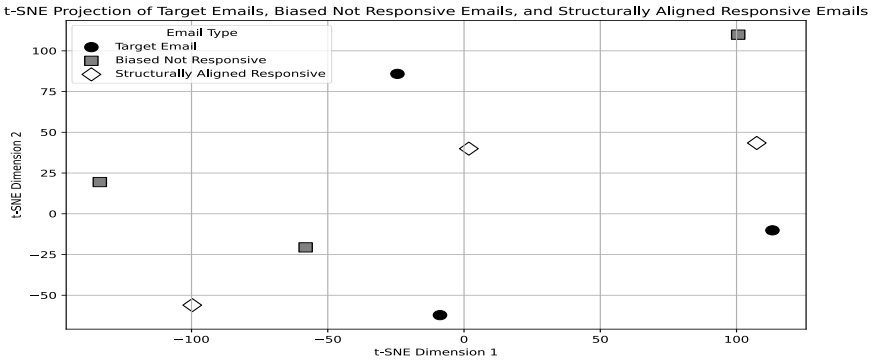

Fig. 1. Distribution of Emails Based on Linguistic Similarities

The t-SNE plot (Figure 1) illustrates the distribution of Target Emails, Biased Not Responsive Emails, and Structurally Aligned Responsive Emails based on their linguistic similarities. The Target Emails are positioned closer to the Structurally Aligned Responsive Emails than to the Biased Not Responsive Emails, suggesting a greater textual similarity.

When Biased Not Responsive Emails, which were similar in structure to the Target Emails, were introduced, the model classified all Target Emails as Not Responsive, despite the presence of Legal Terminology-Rich Responsive Emails. Conversely, when Structurally Aligned Responsive Emails, which shared structural similarities with the Target Emails, were introduced, the model classified all Target Emails as Responsive. These results reinforce the idea that structural similarity played a decisive role in classification for the Fine-Tuned Legal BERT model.

6 Structural Bias and Adversarial Risks

The simulation results highlight the significant impact of dataset composition on model performance. Even minor modifications to the training data can introduce subtle biases, influencing classification outcomes despite the absence of direct replication of the Target Emails. Given that Legal BERT is designed to comprehend not only individual words but also contextual relationships and sentence structures in legal text, an attorney may seek to exploit this capability.

By leveraging Legal BERT's ability to interpret legal language in context, the attorney could lead the model to classify the Target Emails as Not Responsive without detection by the requesting party by using Biased Not Responsive Emails that are structurally similar to the Target Emails but employ different wording. This approach can be framed as a strategic effort to harness the model's advanced understanding of legal language to steer classification results toward avoiding disclosure while mitigating criticism that it constitutes outright manipulation.

Furthermore, in real-world discovery procedures, the requesting party lacks prior knowledge of the content of the Target Emails, making it impossible to identify hidden biases in the training dataset or generate structurally similar Responsive training emails. Therefore, this limitation underscores the necessity of incorporating technical safeguards into discovery protocols to detect and mitigate biases in TAR-based classifications, while ensuring transparency and accountability in the discovery process.

7 Addressing Challenges in TAR Usage Protocols

7.1 Challenges in Implementing Defenses Against Adversarial Attacks

Many previous studies have explored adversarial attacks targeting machine learning vulnerabilities and corresponding defense strategies. For example, Guha et al. [13] identified and analyzed various adversarial attacks targeting machine learning-based TAR models, including biased seed set exploitation, data poisoning, adversarial examples, hidden stratifications, stopping points, and validation methods. However, the defense strategies proposed in such studies remain ineffective unless they are implemented in actual models. Therefore, it is crucial to explore effective methods for integrating these defenses into practical implementations.

If such defense mechanisms cannot be implemented, even the simple attacks demonstrated in the simulations in this study would remain effective and unmitigated, potentially executed undetected by the requesting party. This underscores the necessity of not only proposing technical countermeasures but also ensuring their practical deployment.

In the context of TAR-based discovery, the responsibility for data collection often falls upon the producing party's legal team. This structural dynamic presents challenges in mitigating biases, as the producing party controls the selection and preparation of data used to train TAR systems. While an ideal approach might suggest extensive diversification of training data or the inclusion of external datasets, such measures are

often impractical due to the constrained and case-specific nature of legal datasets. This inherent conflict of interest makes it unlikely that the producing party would voluntarily adopt defensive measures or enhance the system to mitigate these risks.

Furthermore, under U.S. judicial precedent, producing parties typically may select the methodology they use for their TAR process without judicial involvement, provided that it is reasonable (see, e.g., [14]). Producing parties are not required to follow specific search methods dictated by the requesting party (see, e.g., [15]). Absent exceptional circumstances, such as demonstrable unreasonableness, the producing party retains significant discretion in selecting models and algorithms for discovery.

This discretion, combined with various technical vulnerabilities, underscores the urgent need to embed robust safeguards into discovery protocols. Incorporating specific obligations and accountability mechanisms into these protocols is essential to mitigating risks and ensuring a balanced and equitable approach to discovery.

When new defensive technical measures are incorporated into discovery protocols and established as precedents, they are likely to be integrated into future protocols and ultimately become standard practice. Mandating their implementation by the producing party can promote broader acceptance over time, strengthening the overall robustness and fairness of the discovery process.

7.2 Addressing Transparency Challenges in TAR Workflows

Currently, defensive technical measures are not well integrated into TAR workflows, creating significant challenges. Ensuring the robustness and accuracy of TAR models often involves trade-offs between competing factors [16], which must be carefully considered when implementing safeguards. Moreover, the diverse and evolving nature of attacks exploiting model vulnerabilities makes it nearly impossible to predefine all necessary countermeasures within TAR protocols. Addressing these threats cannot be achieved simply by expanding the training dataset or increasing model complexity. Instead, it is crucial to incorporate rule-based safeguards into TAR protocols, such as mutual checks among litigation parties, by enhancing transparency and accountability in the review process.

Given these challenges, transparency in TAR processes is essential for mitigating risks. Requiring the producing party to disclose information about training datasets and algorithms can help prevent manipulation, build trust, and address various adversarial attacks. However, transparency is often limited by intellectual property protections and confidentiality concerns. Moreover, U.S. courts do not necessarily mandate the disclosure of the TAR process or nonresponsive document sets used for training or validation [9], making full transparency difficult to achieve.

Despite these obstacles, transparency remains critical for fostering trust and ensuring compliance, particularly with advanced models like Legal BERT. Effective protocols should incorporate measures such as sharing training data, generating detailed reports, providing standardized guidelines, and enabling traceability of model decisions. Striking a balance between the need for transparency and the challenges posed by the complexity of deep learning models, proprietary algorithm concerns, and substantial

resource requirements is essential to maintaining fairness and reliability in TAR discovery processes.

7.3 Challenges in TAR for Cross-Border Discovery

The cross-border application of U.S. discovery is expanding, increasingly encompassing data stored in foreign jurisdictions. This trend presents significant legal, technical, and ethical challenges, particularly due to the potential exploitation of TAR vulnerabilities to misclassify relevant documents or exclude critical evidence, thereby undermining the discovery process. These challenges are exacerbated by the lack of international consensus on safeguarding TAR, especially when handling sensitive, multilingual, and culturally specific data under conflicting regulations.

Advanced tools like Legal BERT provide enhanced capabilities for processing complex legal texts and multilingual datasets but also introduce risks of intentional misuse. The opacity of AI-driven decision-making processes further complicates these issues, making it difficult to detect and address exploitation effectively.

To mitigate these risks, a comprehensive approach is essential. International standards should include robust protocols for validating TAR algorithms, clear guidelines for cross-border data handling, and mechanisms to detect and respond to misuse. By addressing these vulnerabilities, the global legal community can foster trust, cooperation, and fairness in the use of TAR for cross-border discovery.

8 Conclusion

The simulations in this study demonstrate that adversarial attacks on TAR systems can exclude critical evidence without detection, potentially influencing judicial outcomes. While advanced models like Legal BERT improve legal text processing, they remain sensitive to subtle biases. Defense strategies proposed in computational research are ineffective unless integrated into real-world TAR workflows. As the producing party operates the TAR system, conflicts of interest make voluntary protective measures unrealistic. This underscores the need for not only technical safeguards but also procedural controls in discovery. Therefore, it is crucial to incorporate rule-based safeguards into TAR protocols to enhance transparency and accountability in the review process and strengthen mutual oversight by litigation parties. Transparency is essential for fostering trust and fairness in discovery. While concerns over intellectual property and confidentiality may limit full disclosure, standardized measures—such as validation protocols, independent audits, and mechanisms to detect and address misuse—can significantly enhance the defensibility and integrity of TAR systems. At the international level, harmonization of discovery practices with global data protection laws, such as the GDPR, is also critical to preventing conflicts between legal obligations and ensuring compliance in cross-border cases.

As TAR adoption expands, particularly in cross-border litigation, these safeguards are critical for managing the increasing complexities of legal and technical landscapes. By implementing robust protocols and fostering international collaboration, the global

legal community can modernize discovery processes while ensuring fair and effective resolution of cross-border legal disputes. Strengthening transparency and accountability in TAR-based discovery is essential to advancing equitable access to justice in an increasingly interconnected world.

References

1. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H.: Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023. National Institute of Standards and Technology, Gaithersburg (2024). https://doi.org/10.6028/NIST.AI.100-2e2023
2. Winfield v. City of New York, 2017 U.S. Dist. LEXIS 194413 (S.D.N.Y. 2017).
3. Cavanagh, E.D.: Discovery in federal courts in support of foreign litigation: Lending a helping hand or legal imperialism? Fed. Cts. L. Rev. 13, 81–98 (2021).
4. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 2012 U.S. Dist. LEXIS 23350 (S.D.N.Y. 2012).
5. Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 127, 2015 U.S. Dist. LEXIS 24996 (S.D.N.Y. 2015).
6. Hyles v. New York City, 2016 U.S. Dist. LEXIS 100390 (S.D.N.Y. 2016).
7. Quartararo, M., Poplawski, M., Strayer, A.: The Technology Assisted Review (TAR) Guidelines. EDRAM/Duke Law School (2019). https://edrm.net/resources/technology-assisted-review-tar-guidelines/
8. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The Muppets straight out of Law School. In: Findings of EMNLP 2020, pp. 1–5 (2020). https://doi.org/10.48550/arXiv.2010.02559
9. The Sedona Conference: TAR case law primer, second edition—A project of the Sedona Conference Working Group on Electronic Document Retention and Production [WG1]. Sedona Conference Journal 24(1) (2023). https://thesedonaconference.org/publication/TAR_Case_Law_Primer
10. The Sedona Conference: Commentary on Proportionality in Cross-Border Discovery. 25 SEDONA Conference Journal 669 (2024), https://thesedonaconference.org/publication/Commentary_on_Proportionality_in_Cross-Border_Discovery
11. Zubulake v. UBS Warburg LLC, 220 F.R.D. 212, 2003 U.S. Dist. LEXIS 18771 (S.D.N.Y., Oct. 22, 2003).
12. Agro, M., Aldarmaki, H.: Handling Realistic Label Noise in BERT Text Classification. In: Abbas, M., Freihat, A.A. (eds.) Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), pp. 11–20. Association for Computational Linguistics, Online (2023), https://aclanthology.org/2023.icnlsp-1.2/
13. Guha, N., Henderson, P., Zambrano, D.A.: Vulnerabilities in discovery tech. Harv. J. Law & Tec. 35, 581 (2022). http://dx.doi.org/10.2139/ssrn.4065997
14. Livingston v. City of Chi., 2020 U.S. Dist. LEXIS 160797 (N.D. Ill. 2020).
15. Lawson v. Spirit Aerosystems, 2020 U.S. Dist. LEXIS 64381 (D. Kan. 2020).
16. Tramer, F., Behrmann, J., Carlini, N., Papernot, N., Jacobsen, J.H.: Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In: Daume III, H., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, pp. 9561–9571. PMLR, Virtual Event (2020).

Leveraging LLMs and LegalDocML to extract legal interpretations: a case study on UK legislation and case law

Safia Kanwal[1], Livio Robaldo[1], Joseph Anim[1], Davide Liga[2]

[1] School of Law, Swansea University, UK
{safia.kanwal,livio.robaldo,joseph.anim}@swansea.ac.uk
[2] Department of Computer Science, University of Luxembourg
davide.liga@uni.lu

Abstract. The increasing volume and complexity of legal texts pose challenges in bridging legislative language with judicial interpretation. This paper introduces a novel methodology, along with a corresponding tool, that leverages Large Language Models (LLMs) and the LegalDocML format in a two-phase approach aimed at extracting legal interpretations of UK legislation within UK case law. The UK Publication Office (The National Archives, TNA) is the single institution in the world providing all its legislation and case law in LegalDocML format. Therefore, the tool is not currently applicable to other jurisdictions. Evaluation results demonstrate high accuracy in identifying and extracting key phrases, showcasing the methodology's effectiveness in addressing the diverse contextual meanings of legal language. The tool source code can be accessed through the GitHub repository https://github.com/SafiaK/Odyssey-Terms-Extraction

Keywords: Legal Interpretations, Large Language Models, LegalDocML, Information Extraction

1 Introduction

The law is essential for governance, conflict resolution, and protecting individual rights. However, the exponential growth of legal texts presents challenges in efficiently processing them. A key issue is linking legislative norms to their *legal interpretation* in case law, as judicial decisions define how laws are applied. Courts interpret legislation, establishing precedents, while lawyers analyze case law to craft arguments supporting their clients' positions. Judges ensure consistent legal interpretations, but the time-consuming nature of this process can delay justice and erode public trust.

LegalTech solutions that connect legislation to case law interpretations can streamline legal analysis. Natural Language Processing (NLP) and Large Language Models (LLMs) offer promising approaches, but general-purpose LLMs lag behind domain-specific models. Training datasets capturing legal interpretations

are difficult to create due to the complexity of legal language and evolving judicial decisions [2], [4], [9], [5]. Developing sophisticated methods that understand context, resolve ambiguities, and identify legal relationships is crucial [3].

This paper presents a two-phase methodology using LLMs to extract legal interpretations from legislative texts. Implemented on UK legislation and case law, it first filters legal documents to identify relevant excerpts, then links key phrases from legislation to their case law interpretations. The structured XML format of UK LegalDocML files, prepared by The National Archives (TNA), facilitates this process. Extending the approach to other jurisdictions would require additional modules to process less structured formats, potentially increasing errors.

This research is part of the "Odyssey" project, with TNA playing a key role [3]. The UK is unique in making all primary legislation and jurisprudence accessible in LegalDocML [4], a widely recognized XML standard [6], [7]. By leveraging this structured data, the proposed methodology enhances legal text annotation and lays the foundation for scalable LegalTech applications.

Every UK act may be easily downloaded in LegalDocML from `https://www.legislation.gov.uk` while every case law from 2003 onward may be downloaded in LegalDocML from `https://caselaw.nationalarchives.gov.uk`. The LegalDocML files, which were prepared and validated by a team of human annotators at TNA, clearly structure the legal texts into sections, paragraphs, etc., and contain explicit references between the legal documents. We therefore tailored our prototype to work with the LegalDocML files from TNA, while leveraging the information already present within them.

The rest of the paper is organised as follows: The next section introduces the input data and discusses how phrases from legislation are legally interpreted in case law, i.e., what we aim for our methodology and tool to extract. The methodology is then presented in Section 3, which contains the core of the research presented in this paper. The next section 4 presents the analysis of experiments and evaluation of the results. Section 5 concludes the paper.

2 The input data: the LegalDocML files from The National Archives (TNA)

As explained in the previous section, all UK legislation and case law are publicly available in LegalDocML format through TNA's portals. Each UK act published on the portal `https://www.legislation.gov.uk` can be downloaded in LegalDocML format simply by appending "`data.akn`" (where "`akn`" stands for "Akoma Ntoso") to the end of the URL. For example, the Child Abduction and Custody Act 1985, accessible online via the first URL in (1), can be downloaded in LegalDocML format by following the second URL in (1).

(1) `https://www.legislation.gov.uk/ukpga/1985/60`
`https://www.legislation.gov.uk/ukpga/1985/60/data.akn`

[3] `https://www.nationalarchives.gov.uk`
[4] `https://www.oasis-open.org/committees/legaldocml`

Due to space constraints, we are unable to provide details about the various LegalDocML tags used in the "`data.akn`" file or include substantial excerpts of the act's XML annotation in this paper[5]. However, the XML format is relatively intuitive. By following the second URL in (1), the reader can observe how the XML format neatly organizes legal texts into parts, sections, subsections, etc., each associated with a specific `eId`. The format also explicitly annotates titles (tag `<heading>`), indexes (tag `<num>`), headings, and references (tag `<ref>`), as well as abbreviations (tag `<abbr>`), among other elements. As previously mentioned, the XML structure provided by the LegalDocML format facilitates the straightforward programmatic retrieval of meaningful legal text, a task that would be significantly more labor-intensive in HTML or PDF files.

Similarly, case law can be downloaded in LegalDocML from the portal `https://caselaw.nationalarchives.gov.uk`, but in a different way. At the bottom of each case law web page, e.g., `https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/3257`, a link labeled "Download this judgment as XML" allows users to download the LegalDocML file of the case. However, note that while LegalDocML structures legislative acts into sections (using the `<section>` tag), it structures case law into paragraphs (using the `<paragraph>` tag).

In the LegalDocML files of the acts, certain phrases denoting key concepts within the scope of the act are tagged as `<term>`. For example, the phrase "rights of custody", which denotes a key concept in the Child Abduction and Custody Act 1985, is tagged in the LegalDocML file as follows:

(2)
```
<p>‘‘<term refersTo="#term-rights-of-custody"
eId="term-rights-of-custody">rights of custody</term>’’ shall
include rights relating to the care of the person of the
child and, in particular, the right to determine the child’s
place of residence;</p>
```

Given that they denote key concepts, many of these phrases are legally interpreted in case law. For example, the phrase "rights of custody" is legally interpreted in the case law "[2020] EWHC 3257 (Fam)", which references the act; the fourth paragraph[6] of the case law, for instance, states that in the Mother's opinion the Father's rights of custody were not breached:

(3) *The Mother opposes the Application on the basis:*

(1) That the children's retention in the UK was not in breach of the Father's rights of custody and so, she says, the retention (or their removal) was not "wrongful" within the meaning of Article 3 of the Convention; alternatively

(2) Etc.

[5] The full vocabulary of LegalDocML is available at `http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html`.

[6] `https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/3257#para_4`

Other paragraphs of the case law may include additional legal interpretations (e.g., it is likely that the Father disputes the Mother's interpretation of his rights of custody), as well as the arguments presented by the lawyers to support these interpretations, and ultimately the judges' decision based on the facts and arguments. It is clear, however, that this case law is highly relevant for legal practitioners who must argue similar cases in court, where the question of whether someone's rights of custody have been violated is at issue. The methodology presented in this paper represents the first step towards the creation of an enhanced repository where the links between key phrases in legislative acts and the paragraphs in case law that legally interpret them are made explicit. LegalDocML already includes tags to link acts with relevant jurisprudence[7], which could be utilized to store these connections once identified by the NLP module.

Nevertheless, the example in (2) and (3) is relatively simple. Most relevant phrases from UK legislation legally interpreted in case law are *not* tagged as `<term>` in the LegalDocML file of the act. TNA defined regular expressions to help annotators identify `<term>`s; for instance, rights of custody" in (2) was tagged as a `<term>` because it appears in quotes (. . .'') and is followed by "shall include." However, most key phrases, like the two discussed below, do not follow a fixed pattern linked to an obvious regular expression. Thus, TNA annotators do not tag them as `<term>`, even though they should be, as these phrases are legally interpreted in at least one case law.

Secondly, and more importantly, contrary to the example in (2) and (3), many key phrases are not repeated verbatim in case law, which makes their identification more challenging. The use of LLMs to identify these phrases is therefore highly promising for developing a recommendation system that can suggest potential `<term>`s to TNA annotators with greater coverage and accuracy than regular expressions. LLMs are capable of *paraphrasing* text, enabling them to effectively identify linguistic variants of the target key phrases.

An example is the phrase "physical, emotional and educational needs" from section 1(3)(b) of the Children Act 1989[8]. This phrase is not tagged as `<term>`, like "rights of custody" in the previous example. Still, it is key for the domain of the act as it is legally interpreted in several case law. One of these is "[2024] EWHC 17 (Fam)", specifically its 66th paragraph[9]:

(4) *By contrast to the position of a German court seised of proceedings, whilst the parties have engaged in proceedings in this jurisdiction concerning X's welfare, in the current circumstances, the English court would not have as easy access to the educational and health care professionals engaged with X, and the information concerning his physical, educational and emotional welfare, that will most fully inform the assessment of X's best interests. Etc.*

[7] Specifically, the `<judicial>` tag, see `http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/os-part2-specs_xsd_Element_judicial.html`.
[8] `https://www.legislation.gov.uk/ukpga/1989/41/section/1#section-1-3`
[9] `https://caselaw.nationalarchives.gov.uk/ewhc/fam/2024/17#para_66`

Note that the phrase "physical, emotional, and educational needs" is paraphrased as "physical, educational, and emotional welfare" in this case law. Nevertheless, both expressions refer to the same concept, which is legally interpreted in (4): the judge determined that the information concerning physical, emotional, and educational needs is not easily accessible to the English court.

Identifying links such as the one between "physical, emotional, and educational needs" and (4) using regular expressions would be too difficult, if not impossible. By contrast, LLMs are capable of making these connections, as demonstrated below in this paper.

A third final example is the phrase "controlling or coercive behaviour" from section 1(3)(c) of the Domestic Abuse Act 2021[10], which is legally interpreted in paragraph 115 of "[2023] EWHC 2983 (Fam)"[11]:

(5) *I find that the Father did coerce the Mother into travelling to the UK and signing documents with the effect of fraudulently procuring UK tax credits and that this constituted financial abuse of a controlling and coercive nature. Etc.*

In the judge's opinion, what the Father did can be categorized as "financial abuse of a controlling and coercive nature", which contextualizes the phrase "controlling or coercive behaviour" within the legal discussion of the trial. Once again, LLMs are currently the single available technology capable of recognizing the link between these two excerpts of text.

2.1 Selected case law and corresponding acts

In this paper, we focus on Family Law cases from the past five years, specifically from 2020 to 2024. Running our developed tool on *all* case law from `https://caselaw.nationalarchives.gov.uk` is considered future work. Furthermore, we only considered case law that references at least one UK act. Other cases that reference only case law or other secondary materials are excluded for simplicity, as including them would require the implementation of an additional module to identify which UK acts they (indirectly) reference. Table 1 shows the breakdown of cases containing legislative references by year.

Year	2020	2021	2022	2023	2024	**Total**
N. of cases	26	22	29	80	40	**197**

Table 1. Cases processed per year

As stated earlier, the LegalDocML format structures case law into `<paragraph>`s. These may contain `<subparagraph>`s. However, during initial experimentations,

[10] `https://www.legislation.gov.uk/ukpga/2021/17/section/1#section-1-3`
[11] `https://caselaw.nationalarchives.gov.uk/ewhc/fam/2023/2983#para_115`

we found that <subparagraph>s often lacked sufficient context when analyzed in isolation. Therefore, we selected <paragraph> as the primary unit for data processing in our pipeline to maintain context and ensure accurate analysis. A similar rationale applies to the LegalDocML files of the acts, where <section> was chosen as the primary unit for data processing.

For each <paragraph>, our developed tool determine whether the <paragraph> legally interprets a concept denoted by a phrase occurring within the UK legislation. To this end, as will be explained in the next section, each <paragraph> is linked to a <section> of a UK act referenced in the case law. As mentioned earlier, we only consider case law that references at least one UK act, ensuring that each <paragraph> will be associated with a <section>. Several <paragraph>s also contain explicit references to sections or subsections of a UK act through the LegalDocML tag <ref>. These <ref>s will, of course, be utilized by the module that associates a <section> with each <paragraph>, as the search will be restricted to only those sections mentioned in the <paragraph> (if any).

3 Methodology

The methodology in this research utilizes LLM-based NLP techniques to identify and extract legal terms from UK legislation that are interpreted in case law. Legal interpretation involves clarifying or applying legislative language in specific case contexts. Our approach dynamically identifies interpretations without predefined labels, extracting structured relationships and incrementally building a dataset of pairs (phr, cl), where phr is a phrase from legislation and cl is a <paragraph> from a case law LegalDocML file.

Our methodology consists of two phases. Phase 1 filters <paragraph>s from case law to retain only those likely containing a legal interpretation, pairing them with relevant <section>s of UK acts. Phase 2 extracts the specific phrase from the <section> that is interpreted in the case law.

By leveraging LLMs, this methodology tackles the complexity of legal language and its contextual nuances. It integrates few-shot learning and chain-of-thought [10] reasoning to enhance phrase filtering and extraction tasks.

3.1 Phase 1: matching <paragraph>s from case law with <section>s from UK acts

As explained in the Introduction above, the LegalDocML format eliminates the need to pre-process the input documents, which is required in non-UK jurisdictions where legislation and case law are only available in PDF and HTML formats. From a LegalTech perspective, this provides the UK with a significant advantage over other jurisdictions. As is well known, pre-processing HTML and, even more so, PDF files is highly labour-intensive, which can easily result in an error rate that propagates through the subsequent steps. On the other hand, in LegalDocML files, the text is already structured and easily accessible.

Therefore, the "pre-processing" of our methodology simply involves collecting all `<paragraph>`s from the input case law directly from the LegalDocML files.

The workflow of the first phase is depicted in Figure 1.

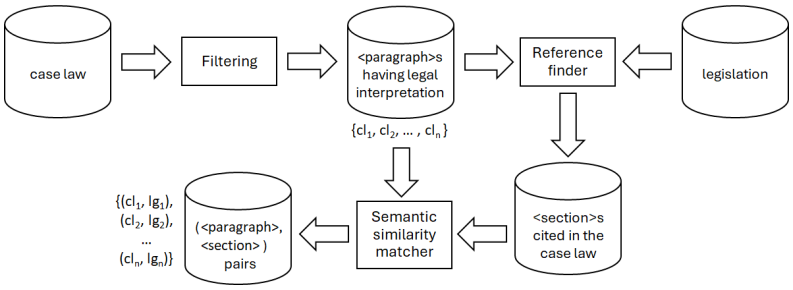

Fig. 1. Workflow for Phase 1 of the proposed methodology

The first step of the methodology, called "Filtering", retains only `<paragraph>`s that convey legal interpretations. Most `paragraph`s in a case contain conventional phrases for formal purposes, background facts, and other non-substantive content. Since these paragraphs are not associated with any section of the UK acts cited in the case, they are excluded a priori.

In order to identify the `<paragraph>`s of interest, we used `GPT-4o mini`. The model was prompted to determine whether the text contains a legal interpretation. Based on our analysis of the input data and the outcomes of our initial experiments, we decided to instruct the model through the following system role:

(6) *You are a legal language model designed to analyze UK case law for paragraphs that contain legal interpretations. Your task is to identify text that interprets or explains legislative terms and concepts.*
- *Accurately identify and analyze any legal interpretations within given texts, focusing on how courts, tribunals, or authoritative bodies explain or clarify the meaning or scope of UK legislation.*
- *Distinguish between mere citations/references and actual legal interpretations. Text that simply cites a statute (e.g., "pursuant to s.100(2)(b)") without any explanatory reasoning or discussion of its meaning does not qualify as interpretation.*
- *Focus on:*
  - *UK legislation (e.g., Acts of Parliament)*
  - *Judicial and statutory interpretation principles (e.g., purposive approach, mischief rule)*
- *Do not consider text as legal interpretation when it:*
  - *Merely mentions the law or quotes statutory wording without explaining it.*

- *Refers to non-UK conventions, treaties, or rulings.*
- *Discusses jurisdictional or procedural issues without interpreting legislative language.*
- *Recites the law verbatim (e.g., "Art. 8 provides...") without additional interpretive commentary.*

The prompt in (6) follows a *chain-of-thought* strategy and incorporates a few-shot learning approach to enhance model performance [1]. This combination enables *in-context learning*, where examples within the prompt guide the model to generate more accurate outputs. By providing demonstrations and step-by-step reasoning, we influence the statistical distribution of responses, improving consistency and accuracy for complex tasks.

To validate the filtering classifier, we manually annotated 12 case laws to create a ground truth dataset. We then applied leave-one-out cross-validation, training the model on 11 case laws while testing on the remaining one, repeating this for all 12 case laws to ensure a robust evaluation.

The validation results, shown in Table 2, indicate high recall but lower precision. This is acceptable since the module primarily reduces the number of `<paragraph>`s processed in later phases rather than aiming for exhaustive identification of legal interpretations. These interpretations are validated by TNA human annotators and, if confirmed, are tagged as `<term>`s in the LegalDocML file and linked to the relevant `<paragraph>`.

Ensuring that most `<paragraph>`s with legal interpretations are retained is more important than minimizing false positives. While high recall may lead to some irrelevant `<paragraph>`s, this is not a significant issue, and increasing the number of examples in the model should further improve results. Table 2 shows that the classifier effectively reduces the number of `<paragraph>`s by 88.27%.

Metric	**Value**
Precision	0.538461538461
Recall	0.913043478260
F1 Score	0.677419354838

Metric	**Value**
Total `<paragraph>`s	20,724
`<paragraph>`s with possible legal interpretations	2,430
% of `<paragraph>`s with possible legal interpretations	11.73%

Table 2. Validation results of the Filtering classifier

Once the set $\{\texttt{cl}_1, \texttt{cl}_2, \ldots, \texttt{cl}_n\}$ of `<paragraph>`s that potentially contain legal interpretations has been identified, all corresponding `<section>`s from the LegalDocML files of the relevant UK acts are retrieved ("Reference finder" in Figure 1). This step is relatively straightforward as it involves following the `<ref>` links provided in the LegalDocML files of the case law. Once again, the use of LegalDocML proves highly advantageous, as the files already include the references that, in the case of plain text processing, would need to be identified automatically, thus introducing a potential error rate that could propagate through

to the final result. The result of this step is a set of <section>s associated with each <paragraph> that possibly conveys a legal interpretation. For those <paragraph>s that include <ref> links to specific <section>s of the acts, only those referenced <section>s are considered. However, only 355 <paragraph>s, i.e., only 14.6% of the 2,430 filtered ones, have an explicit reference to the legislation. <paragraph>s that do not contain any <ref> link are associated with *all* <section>s of any act referenced anywhere in the case law.

The final step of Phase 1 is called the "Semantic Similarity Matcher". The outcome of this step is a list of pairs $\{(\texttt{cl}_1, \texttt{lg}_1), (\texttt{cl}_2, \texttt{lg}_2), \ldots, (\texttt{cl}_n, \texttt{lg}_n)\}$, where $\{\texttt{cl}_1, \texttt{cl}_2, \ldots, \texttt{cl}_n\}$ are the <paragraph>s returned by the Filtering classifier, and $\{\texttt{lg}_1, \texttt{lg}_2, \ldots, \texttt{lg}_n\}$ are the corresponding <section>s that best match the <paragraph>s, among all those associated with each <paragraph> by the Reference Finder. For this step, we employed OpenAI's embedding models[12], specifically `text-embedding-ada-002`. For simplicity, we selected a *single* <section> for each <paragraph>, i.e., the <section> with the highest semantic similarity score, even if multiple acts are referenced in the case.

The resulting set of pairs $\{(\texttt{cl}_1, \texttt{lg}_1), (\texttt{cl}_2, \texttt{lg}_2), \ldots, (\texttt{cl}_n, \texttt{lg}_n)\}$ serves as the input for Phase 2, described in the next subsection. The objective of the subsequent Phase 2 is to identify the specific legislative term (typically a short noun phrase) that occurs verbatim within the <section> $\texttt{lg}_i$ and is legally interpreted in the <paragraph> $\texttt{cl}_i$. These key legislative terms are potential candidates for annotation as new <term> elements in the LegalDocML files, similar to the term "rights of custody" in (2) above.

3.2 Phase 2: Identifying key legislative terms

In Phase 2, the selected pairs (<paragraph>, <section>), where the <paragraph> belongs to a case law and potentially represents a legal interpretation of a key legislative term found in the associated <section>, are processed by an Extractor module. This module employs GPT-4o with *chain-of-thought* techniques. The *chain-of-thought* approach enables the model to break down complex legal reasoning into explicit steps, making its analysis more transparent and reliable.

Therefore, the hypothesis underlying Phase 2 is that, by verbalizing its thought process before reaching conclusions, the model can more effectively identify logical connections between legislative text and its interpretations in case law. The key component of Phase 2's chain of thought are the following:

- **Analysis**: the LLM is tasked with identifying a textual chunk, i.e., a context-providing snippet, within the <paragraph> and aligning it with a corresponding textual chunk in the <section>. The key legislative phrases are then extracted from the selected chunk in the <section> by prompting the LLM to identify the minimal core phrase or phrases. This approach ensures that the system not only extracts key phrases but also considers their interpretation within the given context.

[12] `https://platform.openai.com/docs/guides/embeddings/embedding-models`

- **Extraction Criteria**: phrases are extracted based on semantic equivalence and interpretative value, as not all text within the `<paragraph>` contains the legal interpretation. The system is guided by these extraction criteria:
  - **Textual and Semantic Overlap**: phrases must directly reference or semantically align with the same legal context.
  - **Interpretative Relationships**: extracted chunks should focus on meaningful legal interpretations rather than irrelevant mentions.
  - **Specificity**: key phrases should explain a legal concept rather than a generic concept.
- **Validation and Explainability**: To enhance the system's awareness of its extractions, it is required to provide a confidence score and its reasoning for each extraction. The confidence level ("High", "Medium", or "Low") assigned to each match reflects the degree of alignment between the `<paragraph>` and the `<section>`, showcasing a reference-free evaluation approach. Similarly to the approach in [11], our system assesses the quality and relevance of legal text pairs without relying on reference annotations. While [11] employs question-answering for summary evaluation, we adapt this concept to evaluate the semantic alignment between chunks and phrases within the `<paragraph>` and the `<section>`.
- **Structured Output**: the system is required to structure the output in the following JSON format:

```
{ "case_law_chunk": "text from <paragraph>",
  "legislation_chunk": "text from <section>",
  "key_phrases": ["phrase1","phrase2","phrase3"],
  "reasoning": "reason of extraction",
  "confidence": "level of confidence" }
```

The prompt that implement the above chain of thought is shown in (7):

(7) *You are a specialized legal analyst with expertise in matching legal interpretations between case law and legislation. Follow this systematic process:*

- ***ANALYSIS Phase:***
  - *Identify specific (not overly general) legal concepts or phrases in the case law.*
  - *Find the corresponding, equally specific portion in the legislation. This should be a somewhat longer, context-providing phrase.*
  - *From that longer legislative phrase, also extract the key noun phrase(s) or core concept(s)—the minimal expression that captures the critical legal idea.*
- ***MATCHING CRITERIA:***
  - *Direct textual overlap or near-verbatim references (no paraphrase).*
  - *Semantic equivalence in the same legal context (avoid purely generic wording).*

    - *Clear interpretative relationship (case law explains or applies the legislation).*
    - *Substantive connection (not merely tangential mentions).*
- ***VALIDATION RULES:***
    - *Only extract text that actually appears in each source (verbatim).*
    - *For "`legislation_chunk`", use the longer snippet that captures context.*
    - *For "`key_phrases/concepts`", extract the essential, shorter noun phrase(s) from within that legislation snippet.*
    - *Ensure the match has legal interpretive or explanatory value (avoid trivial or broad phrases).*
- ***OUTPUT STRUCTURE:***
    - *Return a **JSON array** of objects. Each object must contain:*
        * *"`case_law_chunk`": exact phrase from the case law (no rewording).*
        * *"`legislation_chunk`": a longer, context-inclusive phrase from the legislation.*
        * *"`key_phrases/concepts`": the shorter core phrase(s) - verbatim - taken from within "`legislation_chunk`" that most directly capture the legal concept (often a noun phrase).*
        * *"`reasoning`": brief explanation of how the term interprets/applies to the legislation.*
        * *"`confidence`": "`High`", "`Medium`", or "`Low`" based on how closely they match in legal meaning.*
- ***RULES:***
    - *Extract only exact phrases from source texts.*
    - *No rephrasing or inference.*
    - *Include only paired matches with clear legal interpretation.*
    - *Return raw JSON without formatting or explanation.*

A distinctive feature of our approach is its emphasis on explicit reasoning by the LLM. By requiring explanations and encouraging the model to "think aloud", chain-of-thought prompting has been shown to enhance performance [8].

Tables 3, 4, and 5 show case law chunks that legally interpret the key legislative terms "child's welfare", "rights of custody", and "controlling or coercive behaviour", which appear in the Children Act (1989), the Child Abduction and Custody Act 1985, and the Domestic Abuse Act 2021 of UK legislation, respectively. These tables show how the same legal concept can be expressed in different forms within the legal narrative.

The first column of the table contains the year, the case law ID, and the paragraph ID (pId) within the case law. For example, '2020/877-10' refers to paragraph 10 in [2020] EWHC 877 (Fam)[13].

[13] `https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/877#para_10`

Year/Id-pId	case_law_chunk
2020/877-10	welfare of the child, while a primary consideration, is not the paramount consideration
2020/3496-10	J's best interests
2020/2878-141	welfare analysis itself involves a balance of interference with and promotion of her rights
2024/1156-44	the judge must consider the child's welfare now, throughout the remainder of the child's minority and into and through adulthood
2021/33-82	adoption was the only realistic option for this child
2021/2931-124	welfare questions in circumstances where moving the child by reason of an unacceptable delay in securing registration may conflict with the child's wider welfare needs

Table 3. child's welfare (Children Act 1989)

Year/Id-pId	case_law_chunk
2020/1599-86	the child herself objects to being returned
2020/3257-113	breach of the Father's rights of custody
2020/1903-59	removal was indeed in breach of the mother's rights of custody
2022/1827-32	father did not, at the time P was removed from the Republic of Ireland, have rights of custody
2024/1282-14	judge considering a return order
2023/2082-100	the exercise of the discretion under the Convention

Table 4. rights of custody (Child Abduction and Custody Act 1985

Year/Id-pId	case_law_chunk
2022/2755 - 9	controlling, coercive or threatening behaviour, violence or abuse
2023/2983 - 115	financial abuse of a controlling and coercive nature
2023/505 - 44	cutting her off from friends and family

Table 5. controlling or coercive behaviour (Domestic Abuse Act 2021)

4 Analysis and Evaluation

The evaluation of our methodology primarily focused on assessing the accuracy and reliability of extracting the final key terms, such as "child's welfare", rights of custody", and "controlling or coercive behaviour". This section presents the analysis of the results along with their evaluation, including some limitations of our work, which set the basis for future improvements.

4.1 Analysis of the results

Of the 2,430 <paragraph>s processed in Phase 1, our system successfully extracted one or more key phrases from 2,066 <paragraph>s, accounting for 85% of the total. Each extracted phrase was linked to the JSON template shown above in "Structured Output."

Conversely, for 364 <paragraph>s, the system either failed to associate a <section> with the <paragraph> or did not identify any key phrases within the text. This occurred either because the <paragraph> did not contain any legal interpretation (recall that the precision of Phase 1 is 0.54%), or because the interpreted content does not appear in UK legislation. On a deeper analysis, we came to know that 308 paragraphs actually do not have any legal interpretation. One of the other reasons we found that other legal documents, such as the Hague Convention or the European Convention on Human Rights, are frequently referenced in case law and could be subject to interpretation. However, since our study focuses exclusively on UK legislation and we extract only <section>s from UK laws, the system was unable to process these <paragraph>s.

3008 key legislative terms were extracted from 2066 <paragraph>s, with several <paragraph>s yielding more than one key term, as explained earlier. These terms occur verbatim in the <section> associated with each <paragraph>. Most extracted key phrases are short noun phrases, averaging 4.14 words. 49 legislative acts were mentioned in the selected <paragraph>s, with the most frequently mentioned being the Children Act 1989.

Finally, the system assigned confidence levels ("High", "Medium", "Low") to its outputs based on the semantic alignment between case law and legislative text. Of the extracted key terms, 74.1% (2256 terms) were classified as "High" confidence, 25.8% (747 terms) as "Medium" confidence, and only 0.2% (5 terms) as "Low" confidence. These results underscore the system's confidence in identifying meaningful connections.

4.2 Evaluation

The quality and relevance of key phrases were assessed under the supervision of a legal expert with a PhD in law, ensuring a high standard of evaluation. Given the complexity and nuanced nature of legal language, the expert's review was indispensable in ensuring both accuracy and consistency with established legal principles. The evaluation was based on two key criteria: the relevance of each key phrase to the specific case law described in the <paragraph>, and its validity as a recognised legal concept within the context of the act to which the <section> associated with the <paragraph> pertains.

The expert review process involved multiple steps. First, the extracted key phrases, along with their associated <paragraph>s, reasoning, and contextual details, were compiled into an organised spreadsheet. The spreadsheet was structured to enable a systematic evaluation and included two dedicated columns (`key_phrases_check` and `reasoning_check`) with drop-down options of "yes" or "no" for streamlined assessment. Specifically, the legal expert was tasked not

only with evaluating whether the key phrase extracted from the <section> was indeed legally interpreted in the <paragraph>, but also with assessing whether the reasoning provided by the LLM about *why* the <paragraph> conveyed a legal interpretation of that key phrase was sound. We consider the evaluation of the latter to be even more critical than the former, as it assesses the *explainability* of our results and lays the groundwork for characterizing, in future research, different sub-categories of legal interpretations.

This spreadsheet was then provided to the expert for annotation and validation. The expert was instructed *not* to seek additional information from TNA's portals or any external sources to ensure that their evaluation was based solely on the information available to the LLM for each <paragraph>-<section> pair.

The legal expert spent approximately four weeks conducting an exhaustive review of the spreadsheet. As explained above, the expert not only verified the accuracy of the extracted phrases but also scrutinised the underlying reasoning, ensuring that the logical connections drawn between the key phrases and their legal implications were sound and well-supported by established legal principles.

As shown in Table 6 on the left, the legal expert marked most spreadsheet cells as "yes," confirming the LLM's ability to identify and explain legal interpretations of key terms from UK legislation within case law. Notably, the model achieves a reasoning accuracy of 98.29% when highly confident, demonstrating its capability to handle complex legal interpretation tasks effectively.

Table 6 on the right presents the legal expert's analysis of the 364 <paragraph>s the system discarded, either due to a failure to associate a <section> with the <paragraph> or because no key phrases were identified. According to the expert, 84.6% of these paragraphs do not contain any legal interpretation, while 8.52% include one but not of key phrases in UK legislation. In both cases, the system correctly discharged them. Only 6.8% of <paragraph>s were mistakenly discharged—i.e., the system failed to recognize either their legal interpretation or the specific key phrase being interpreted.

Confidence	**Key Phrase Accuracy**	**Reasoning Accuracy**
Low	100%	20%
Medium	99%	32%
High	99.60%	98.29%

Reason discharged <paragraph>	**%**
Paragraphs do not have legal interpretation	84.6%
Interpretation is not of UK legislation	8.52%
Have legal interpretation but system failure	6.8%

Table 6. Analysis of Accuracy and Failure Reasons

In addition, when consulted on the legal soundness of our overall research endeavour, the expert noted that, while the extracted key phrases were interpreted accurately within their respective <paragraph>, this approach may not capture the full complexity of legal reasoning. Legal analysis is inherently multifaceted, often requiring the simultaneous interpretation of multiple legislative <section>s. A single <paragraph> in a case law document may encompass legal concepts influ-

enced by several <section>s, as practitioners frequently consider the combined effect of different provisions to construct arguments or derive conclusions.

This interconnected nature of legal interpretation poses a significant challenge for the methodology, which we aim to address in future work. The current version relies on mapping a <paragraph> to a single corresponding <section>. By restricting key phrases to one <section>, the analysis may overlook nuances from cross-references and interdependencies in the legal text. The expert emphasized that this limitation is a critical factor that could impact the depth and comprehensiveness of extracted legal interpretations.

The second key observation highlighted by the legal expert was the challenge posed by very short <paragraph>s. In some cases, these <paragraph>s lacked sufficient context to enable a complete legal interpretation, necessitating a reference to preceding <paragraph>s within the case law document. This reliance on preceding text underscores the contextual nature of legal language, where meaning often emerges from earlier arguments or explanations. In the current version of our methodology, a consistent unit of processing was defined, with <paragraph>s chosen as the standard unit for analysis. While this approach effectively addresses the majority of <paragraph>s, it does present limitations in situations where context is fragmented across multiple <paragraph>s. However, the expert noted that such cases were relatively rare and did not significantly impact the overall efficacy of the process. For the purposes of this study, this trade-off was considered acceptable. However, it highlights an area for refinement in future iterations of the methodology, where the LLM should be enabled to also examine the <paragraph>s that precede the one under analysis.

5 Conclusions

This paper presented a two-phase methodology for linking legislative text excerpts to their legal interpretations in case law, combining Large Language Models (LLMs) with structured legal data (LegalDocML). Implemented on UK legislation and case law from The National Archives (TNA), the approach first identified and filtered <paragraph>s conveying legal interpretations and associated them with relevant <section>s. In the second phase, it extracted key phrases occurring in these sections and linked them to their legal interpretations.

The motivation for this work lies in improving the annotation of legal terms in UK legislation. Current methods rely on rigid regular expressions, which struggle to capture linguistic variations and contextual nuances. By leveraging LLMs' paraphrasing capabilities, our approach identifies diverse expressions of legal concepts and connects them to their case law interpretations more effectively than regex-based methods. This marks a step toward automated systems that enhance legal text analysis and understanding.

Beyond improving legal annotation, this methodology has significant implications for LegalTech. It enables applications that assist lawyers in case preparation and judges in harmonizing legal interpretations, ultimately making legal analysis more efficient and contributing to fairer, more consistent legal decision-making.

The results demonstrated strong performance, highlighting the synergy between LLMs and LegalDocML data. While promising, the approach can be further improved. Future work will expand coverage to additional legal domains, refine section mapping, and develop a validation tool for TNA annotators to enhance LegalDocML annotations. Over time, the growing dataset could support training domain-specific LLMs, further improving precision and scalability.

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), `https://arxiv.org/abs/2005.14165`
2. Jayakumar, T., Farooqui, F., Farooqui, L.: Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. In: Preoțiuc-Pietro, D.e.a. (ed.) Proc. of the Natural Legal Language Processing Workshop 2023. Association for Computational Linguistics (2023)
3. Katz, D.M., Bommarito, M.J.: Measuring the complexity of the law: The united states code. Artificial Intelligence and Law **22**(4), 337–374 (2014)
4. Liga, D., Robaldo, L.: Fine-tuning GPT-3 for legal rule classification. Computer Law & Security Review **51**, 105864 (2023)
5. Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D.: MultiLegalPile: A 689GB multilingual legal corpus. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2024)
6. Palmirani, M.: Legislative Change Management with Akoma-Ntoso. Springer Netherlands, Dordrecht (2011)
7. Palmirani, M., Vitali, F.: Akoma Ntoso for Legal Documents, pp. 75–100. Springer Netherlands, Dordrecht (2011)
8. Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., Chen, H.: Reasoning with language model prompting: A survey (2023), `https://arxiv.org/abs/2212.09597`
9. Stern, R., Rasiah, V., Matoshi, V., Bose, S.B., Stürmer, M., Chalkidis, I., et al.: One law, many languages: Benchmarking multilingual legal reasoning for judicial support (2024), `https://arxiv.org/abs/2306.09237`
10. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. CoRR **abs/2201.11903** (2022), `https://arxiv.org/abs/2201.11903`
11. Xu, H., Ashley, K.: A Question-Answering Approach to Evaluating Legal Summaries. IOS Press (Dec 2023). https://doi.org/10.3233/faia230977, `http://dx.doi.org/10.3233/FAIA230977`

Legitimacy Justification and Legal Regulation: Platform-Based Prior Review of Copyright in Fast Dramas under Advancing Algorithmic Technologies

Shaowei Ji[1*][0009-0001-1694-2823]

[1*]School of Political Science and Law, University of Jinan, Jinan 250004, China

[1*]16626990180@163.com

Abstract. With the rapid development of fast dramas, the issue of infringement has become increasingly prevalent. Whether the 'safe harbor' principle for platform liability applies in the context of advanced algorithmic technologies has become a contentious focal point in both academic and practical circles. First, the research analyzes the controversy surrounding this issue in both academic and practical circles, and then, in light of the current development of algorithmic technologies, points out the limitations of the 'safe harbor' principle in addressing copyright infringement issues in fast dramas. Further, analyze the technical feasibility of platform copyright pre-screening in light of the development of algorithmic technologies, as well as the legal and societal foundations that support it. Finally, determine the scope of platform's pre-screening content for Fast Dramas' copyright, the boundaries of liability for copyright infringement in Fast Dramas, and propose corresponding improvements in legal regulations. It aims to promote a balance of rights between platforms and creators, achieve regularized regulation of the Fast Drama industry, and advance the development of intellectual property protection.

Keywords: Fast drama, Copyright, Platform Liability, Algorithmic Content Moderation

1 Introduction

With the rapid development of the short video and Fast Drama markets, Fast Dramas have become an important medium for attracting users and traffic. According to the 'China Fast Drama Industry Development White Paper (2024)', as of June 2024, the user base of Fast Dramas in China has reached 576 million, accounting for 52.4% of the total internet users [1]. The "2023-2024 China Fast Drama Market Research Report" indicates that in 2023, the market size of Fast Dramas in China exceeded 37.39 billion yuan, representing a 267.65% growth compared to 2022. It is expected that by 2027, the market size of Fast Dramas will exceed 100 billion yuan [2]. Fast Dramas, with their concise and impactful format and convenient dissemination characteristics, have rapidly occupied the traffic gateways of platforms, becoming an important source of revenue for content creators and platforms. However, with the rise of Fast Dramas, issues of copyright infringement have also followed closely behind. The "2024 Fast Drama Anti-Infringement and Anti-Piracy Action Plan" shows that in 2024, the WeChat platform handled 163 instances of Fast Drama copyright infringement across 82 mini-programs, while Douyin has processed over 78 million infringing pieces of content since the beginning of 2024. Pirate groups steal legitimate Fast Drama works through illegal means, edit video clips, and sell them on social media platforms, creating a complete illicit industry chain that severely impacts the healthy development of the Fast Drama industry. The 2023 "Short Drama Copyright Protection Report" reveals that among the 331 short dramas monitored, each work faced an average of 12,224 infringement links, totaling up to 405,300 links. The economic losses suffered by copyright holders are significant, highlighting the urgent need to address copyright protection issues within the industry [3].

In the current copyright infringement review framework, short video platforms typically rely on the "Red Flag Rule" to define responsibility. However, there is still ambiguity regarding when the "red flag" should be raised [4]. Especially in today's era of advanced algorithmic technology, platforms possess unprecedented content recommendation capabilities, allowing them to accurately identify user preferences and push relevant content. This raises the possibility that platforms, after becoming aware of infringing content, may intentionally allow algorithmic recommendations to spread infringing videos for profit, only to later deny knowledge of the infringement and invoke the "Safe Harbor Rule" to evade responsibility [5]. In the current technological environment, platforms are fully capable of proactively identifying and filtering potential

infringing content through their algorithmic systems, rather than solely relying on user reports or notifications from copyright holders. The technological tools and data analysis capabilities of platforms are no longer limited to being a neutral information provider; they have become active content distributors and, as such, should bear corresponding responsibility for the legality of the content [6].

2 The Controversy Over Platforms' Pre-Screening Obligations

In 2018, iQIYI Inc. filed a lawsuit against ByteDance Ltd., alleging that the latter's subsidiary, Toutiao, had broadcast user-uploaded short videos of Story of Yanxi Palace, infringing iQIYI's right of information network dissemination. After four years of litigation, in December 2021, the first-instance court found ByteDance to have assisted in the infringement and ordered it to pay iQIYI 2 million RMB. This case has been dubbed the first "algorithm recommendation infringement case." Although the case was ultimately settled on appeal, the key dispute remains unresolved: whether short video platforms, by using algorithmic recommendation technologies, should bear a higher duty of care in relation to user infringement behaviors [7].

Some scholars argue that platforms should not be burdened with excessive obligations, as increasing their duty of care could lead to an overexpansion of platform censorship powers and a failure of judicial order [7]. Adopting the view that "the greater the benefit, the greater the risk, the stronger the responsibility" to determine a platform's liability for algorithm usage is overly simplistic. This is because algorithmic recommendation technologies do not directly involve specific content but merely provide users with content links that meet their needs, rather than the content itself being aimed at the general public [8]. In the case of Tencent Video v. Douyin over the copyright infringement of North-South Still Believes in Love, the court held that the platform had no statutory obligation for "pre-screening and filtering." Even if infringing content appeared on the platform, it did not automatically imply platform liability [9].

However, there are opinions consistent with the first-instance court's stance in the first algorithm recommendation case. These argue that as an online intermediary service provider, the platform should bear an examination obligation, which aligns with the basic requirements of the objective value order theory. Furthermore, establishing such an obligation would raise intellectual property liability to a level commensurate with economic and social development [10]. As algorithmic recommendation technology has become more widespread, platforms are no longer passive information

providers but rather active content "distributors" who shape content dissemination. Therefore, imposing stricter standards of care on platforms is feasible [11]. In the case of Guangzhou Lizhi Network Technology Co., Ltd. v. Jiecheng Huashi Wangju (Beijing) Cultural Media Co., Ltd., the court held that the platform's algorithmic recommendation services substantially enhanced the efficiency and scope of infringement dissemination. Thus, the platform should bear a higher duty of care for user infringing behaviors [12].

Currently, there remains considerable debate in both academic and practical circles regarding whether platforms should be held responsible for copyright review obligations arising from algorithmic recommendations. Scholars and courts have varying interpretations and stances on this issue.

3 Justification for Platform's Pre-emptive Review Obligation in Fast Dramas Copyright Protection

With the development of algorithmic technology, the 'safe harbor' rule has become increasingly limited in the protection of Fast Dramas' copyrights. The platform's responsibility in Fast Drama copyright infringement cases is becoming more prominent, especially in the pre-emptive copyright review process. In the protection of Fast Drama copyrights, pre-emptive review by platforms is already technically feasible and has a certain legal and social foundation.

3.1 Limitations of the "Safe Harbor" Rule in Fast Dramas Copyright Protection

In the information network, online service providers fulfill their obligations under the 'Notice-and-Takedown' rule as stipulated in the Civil Code and related laws, and Fast Drama infringement also follows this rule. However, the passivity of this rule makes it difficult for short video platforms to cope with the surge in short video infringement cases [13]. Under this rule, platforms bear a passive duty of care. Especially with the rapid development of various short videos and the surge in Fast Drama uploads, if platforms continue to handle matters according to the "Notice-and-Takedown" rule, allowing users to upload Fast Drama content to their servers for others to stream, this creates a mismatch between the platform's earnings and its responsibilities. This leads to an imbalance of rights and obligations. On the other hand, in the context of an

overwhelming number of infringement notices (where platforms receive a large volume of mixed, false, and fraudulent notices and are overwhelmed in dealing with them), it also greatly reduces the platform's processing efficiency [14]. Copyright holders face the challenge of sending infringement notices across numerous platforms and websites with vast amounts of video content, which is time-consuming and labor-intensive [15], making it difficult to protect their legal rights. Especially for works with strong time sensitivity, the copyright holder's losses are difficult to recover [16]. The "Notice-and-Takedown" rule cannot effectively prevent infringement, including those related to Fast Dramas. The "Notice-and-Takedown" rule is essentially designed to protect platforms from directly bearing infringement liability due to user actions, and therefore does not require platforms to proactively review the copyright issues of uploaded content. As a result, no matter how much the rule is modified or improved, it cannot effectively curb the surge in Fast Drama copyright infringement disputes.

3.2 Technical Feasibility of Pre-Upload Copyright Review for Fast Dramas

As early as 2007, YouTube invested in the development of Content ID to address content copyright issues. The system has undergone multiple improvements, significantly enhancing its content monitoring capabilities. For example, it can use hash algorithms to tag potentially infringing videos and prevent these videos from being uploaded again. In the fourth quarter of 2017, YouTube removed over 10 million illegal videos, with 6.7 million of them being automatically flagged by monitoring software. Approximately 75% of the infringing videos were taken down before users could view them. Although this technology still has some instances of false positives and missed detections, its effectiveness in copyright protection cannot be overlooked.

Similarly, Facebook and TikTok also place great emphasis on handling copyright disputes, focusing on content moderation, and gradually leveraging artificial intelligence to strengthen content review. In 2018, Facebook hired over 20,000 people to assist with content moderation. Although this number may seem large, it is reasonable given the platform's vast user base and the massive amount of content. In 2021, TikTok successfully blocked over 200 million violating videos through its AI review system. Compared to the initial method that relied entirely on manual review, the workload for human moderators was reduced by nearly 70%. According to a report by Sina Technology, as of 2021, the accuracy rate of AI content moderation generally hovers around 90%. Tencent Cloud's intelligent content moderation has an accuracy rate of 98.5%,

while Baidu's intelligent content moderation improved its accuracy rate by 20% between 2020 and 2021, leveraging big data and machine learning technologies.

Although content recognition technology still experiences some false positives and missed detections, significant breakthroughs have been made in technological development compared to a few years ago. Advances in deep learning and big data technologies have greatly improved the efficiency and accuracy of content moderation.

In 2021, the performance of training models such as BERT had already surpassed human benchmarks. This technological advancement enabled AI models to more accurately identify and filter sensitive information, reducing the burden on human moderators. In the field of computer vision, breakthroughs in algorithms such as convolutional neural networks (CNNs) have also enhanced the ability to recognize images and video content more effectively. The number of companies in China's AI content moderation industry has been increasing year by year, including Tencent, Alibaba, and Baidu Cloud. Each of these companies has also introduced AI moderation solutions, such as Alibaba's Cloud Shield and Tencent's Tianyu Risk Control Platform. The continuous advancement of these technologies demonstrates the enormous potential of AI in large-scale content moderation and proves that platforms are fully capable of implementing preemptive copyright review for Fast Dramas.

Moreover, Fast Dramas are just one type among the many categories of short videos, but they have distinct characteristics compared to other short videos. Platforms can utilize technologies such as hash algorithms, deep learning models, and computer vision to tag and analyze content. Combined with a pre-release filing system, it is technically feasible to implement preemptive copyright review for Fast Dramas.

3.3 The Legal Basis for the Preemptive Copyright Review of Fast drama

The pre-authorization copyright review of fast dramas has a solid legal foundation. First, the Copyright Law of the People's Republic of China was amended in 2001 to establish the right of information network communication, which clearly stipulates that no entity shall infringe upon others' information network communication rights, and violators will bear corresponding legal liabilities. Subsequently, the Regulations on the Protection of the Right of Information Network Communication (2006) further defined the behavior of distributing infringing content through information networks without authorization, specifying the legal responsibilities of internet platforms in preventing infringement.

At the platform level, the Regulations on the Administration of Internet Audio-Visual Program Services stipulate that short video platforms must conduct pre-authorization content review to ensure that the published works comply with relevant laws and regulations, particularly concerning political, moral, and other issues. The Network Short Video Platform Management Norms issued in 2019 further refined the content review responsibilities of platforms, emphasizing the need for a "pre-review-before-broadcast" system, and the establishment of a qualified review team to ensure the professionalism of the review process.

In addition, national policies on the copyright responsibilities of short video platforms have been increasingly reinforced in recent years. The New Generation Artificial Intelligence Development Plan (2017) explicitly calls for the application of AI technology to enhance the efficiency and accuracy of internet content review. In 2022, the State Administration of Radio and Television issued a notice further emphasizing the platform's review obligations and requiring enhanced content supervision of individual creators and unregistered entities.

Internationally, the copyright review responsibilities of platforms have also gradually increased. The Digital Millennium Copyright Act (DMCA) in the United States originally provided platforms with a "safe harbor" provision for liability exemption [8]. However, with the diversification of platform functions, increasing judicial practices require platforms to take on greater responsibility for user-uploaded content. In 2019, the European Union adopted the Digital Single Market Copyright Directive, which explicitly requires online platforms to take effective measures to prevent the distribution of unauthorized content and to swiftly remove or block infringing content upon receiving a notice, thus shifting to a stricter "prevention-governance" model [17]. South Korea has also enacted the Telecommunications Business Act and the Internet Content Filtering Ordinance to enhance the supervision and management of internet content.

In conclusion, both domestically and internationally, platforms' pre-authorization review responsibilities have become an inevitable trend in copyright protection. From the perspective of intellectual property protection, platforms are gradually shifting from traditional neutral content providers to information governors, now bearing more stringent copyright compliance responsibilities. With the advancement of technology, particularly AI content review systems, platforms will be more efficient in fulfilling their review obligations, but they must also balance copyright protection with freedom of expression and innovation development, ensuring a harmonious alignment between legal responsibilities and market needs.

3.4 The Social Basis for the Preemptive Copyright Review of Fast drama

The establishment of copyright review obligations for short video platforms is not only a reflection of the legislator's subjective intentions but also a justification based on the realities of social relationships. As a new legal obligation, its rationality and necessity need to be deeply rooted in external social facts and the foundational social relationships. This external foundation is mainly reflected in the following two aspects.

3.4.1 The Rapid Growth of the Short Video Industry and the Escalation of Copyright Issues.

The fast drama industry has grown rapidly since 2021, with its market size increasing from 3.68 billion yuan to 37.39 billion yuan by 2023 [18], showing explosive growth in just two years. Numerous platforms have followed this trend by launching dedicated short drama modules. For example, Dianzhong Technology operates the "Hippo Theater" and "Fanhua Theater," while platforms like Douyin, Kuaishou, Taobao, and Pinduoduo have added short drama sections [19]. Dedicated short drama platforms like "Tomato Short Drama," "Bai Chuan Chinese," and "Malt Short Drama" have also emerged.

Through fast dramas, platforms not only attract a large number of users but also achieve revenue growth through diversified profit models such as paid content, membership services, and advertising placement [20]. For instance, Kuaishou's total revenue in the third quarter of 2023 was 27.948 billion yuan, representing a 20.84% year-on-year increase, with paid short drama revenue growing by more than 300%. Tencent has also used fast dramas to boost its online advertising business. However, the industry's growth has been accompanied by increasingly severe copyright issues, including clipping, re-uploading, rewriting, and IP infringement, all of which disrupt the copyright market order.

3.4.2 Imbalance in the Allocation of Copyright-Related Rights and Obligations.

Under the existing legal framework, copyright holders can only demand legal liability from infringers if they can detect the infringement and provide sufficient evidence. The existing legal framework clearly shifts the burden of review onto copyright holders. However, in the era of information explosion, with vast amounts of content across multiple platforms, it is extremely difficult for rights holders to detect in-fringement in a timely and effective manner. The limitations of the current legal system reveal that

relying solely on rights holders to detect infringement and pursue accountability is no longer suitable for the speed and scale of modern information dissemination.

Internet-related activities conducted by platforms should fall under legal regulation. Like other societal activities, they need to balance the rights and obligations of different parties. Although the existing legal framework does not explicitly require platforms to conduct prior copyright reviews, this does not mean that this relationship of rights and obligations should not be legally regulated and adjusted. Any social fact, whether long-standing or newly emerging due to societal developments, that can objectively influence people's rights and obligations should be subject to legal regulation. The current competitive structure of platforms weakens the role of market regulation, and the negative externalities of the platform economy make it clear that fast drama infringement issues cannot be solved purely through market forces. Without legal intervention, the imbalance in the rights and obligations between platforms and copyright holders will continue to worsen, destabilizing the internet society and hindering the long-term development of intellectual property protection and cyberspace.

4 The Scope of Preemptive Copyright Review for Fast drama and the Limits of Platform Liability

To achieve prior review of fast drama copyrights by platforms, it is essential to define the scope of content to be reviewed and clarify the boundaries of responsibility in infringement cases. At the same time, to ensure more effective copyright protection, the existing legal regulations should be improved, and efforts should be made to establish a comprehensive copyright protection mechanism.

4.1 The Scope of Preemptive Copyright Review for Fast drama

In the pre-examination of fast drama copyright, platforms should comprehensively apply advanced technologies such as artificial intelligence, blockchain, hash fingerprinting, and content recognition to implement end-to-end, intelligent, and precise copyright compliance management. This includes originality and authorization review, legal content source verification, copyright marking and management, and monitoring user uploads and editing activities. In the originality and authorization review process, platforms can use hash fingerprint comparison and deep learning technology to establish an efficient content similarity detection system. By combining knowledge graphs

and semantic analysis, platforms can accurately identify the similarity between uploaded content and existing works in the database in terms of text, visuals, and editing styles, thus determining whether the content constitutes plagiarism or unauthorized adaptation. For adapted works, platforms can integrate smart contracts and blockchain-based proof mechanisms. When users upload works, the platform can automatically verify their authorization status, ensuring the existence of legitimate usage or adaptation rights, and preventing the risks of forged or duplicate authorizations. Regarding the legality of content sources, platforms can utilize computer vision, audio and video fingerprinting, and digital watermarking technologies to automatically analyze multimodal data in fast dramas, identifying whether they contain unauthorized third-party materials (such as music, images, or video clips). In combination with dynamic rights management systems, platforms can set up automatic authorization matching and alert functions for materials, ensuring that content compliance is checked when uploaded and offering embedded automatic authorization or copyright trading interfaces to enhance compliance efficiency. Additionally, platforms should establish smart contract-driven copyright registration and tracking systems to ensure clear and traceable copyright information for each fast drama (such as rights holders, authorization methods, usage scope, etc.). By drawing on global mainstream content management systems like Content ID, platforms can generate hash codes (SHA-256) and distributed proofs (IPFS + blockchain) for verified works, enabling comprehensive tracking and comparison of copyright status across the network, preventing unauthorized re-uploads, secondary distribution, and hidden infringement activities. To monitor user uploads and editing behaviors, platforms can develop deep learning and big data-driven infringement risk prediction models, analyzing user editing techniques, material splicing patterns, and audio-video processing features to automatically detect potential infringement behaviors. When the risk level is high, automatic blocking, manual re-examination, and smart appeal mechanisms can be triggered to ensure timely identification and processing of infringements. For repeat infringers, platforms can establish cross-platform user credit scoring systems, improving infringement traceability through behavioral trajectory analysis and blacklist sharing mechanisms, thereby reducing the likelihood of evading scrutiny. The intelligent governance framework for copyright review in fast dramas, combining technological innovation with legal safeguards, not only improves the efficiency and accuracy of copyright audits but also effectively prevents infringement, driving the healthy development of the fast drama industry within a framework of technology-driven progress and legal protection.

4.2 Boundaries of Platform Liability in Fast drama Copyright Infringement

In fast drama copyright infringement cases, the key to determining platform liability lies in whether it has fulfilled its reasonable review obligations and effectively identified and prevented infringement based on its technical capabilities and content dissemination characteristics. Platforms face two types of responsibility in content review: first, failure to fulfill the prior copyright review obligation, and second, despite fulfilling the review obligation, failure to effectively identify infringing content, resulting in the spread of the infringement.

Firstly, in cases where the platform fails to fulfill its prior review obligation, if the platform does not employ sufficient technical means or measures to conduct pre-upload copyright review, leading to the uploading and widespread dissemination of infringing content, the platform should bear copyright liability in accordance with the law. Secondly, even if the platform has conducted a prior review, if it fails to identify potential infringement in a timely manner—especially when content reaches a "should-know" threshold and spreads rapidly among users—the platform should initiate a secondary review mechanism to prevent further infringement. For example, when the interaction of a fast drama reaches a set threshold on the platform (such as shares, views, etc.), the platform should treat this content as high-risk and immediately carry out a more in-depth review.

In the realm of fast dramas, where content updates frequently and audience interaction is high, the platform's responsibility boundaries are particularly ambiguous, especially when facing infringement behavior that bypasses review through technical means (such as web scraping, splicing, modification, etc.). This type of infringement typically arises from limitations in the platform's technical capabilities. The determination of platform responsibility should not only consider the technical protective measures already taken but also the platform's recommendation algorithm's operating principles and content dissemination features. For example, the platform's recommendation algorithm often pushes content to prominent positions, such as trending topics or homepages. In these cases, the platform should assess whether the content has already spread widely based on the level of interaction, thus triggering a secondary review mechanism.

In response to the platform's technological limitations, legislation could establish a "reasonable review obligation" standard, requiring platforms, based on their size, technical capacity, and economic capability, to implement basic copyright protection measures, such as hash fingerprinting and content recognition technologies. The

standard should also raise the infringement compensation liability in cases where the platform fails to fulfill this reasonable review obligation. This "reasonable review obligation" standard aims to balance the platform's technical capabilities with its responsibility scope, preventing negligence due to cost and business pressures.

At the same time, the government can introduce incentive policies, such as tax reductions and financial subsidies, to encourage platforms to invest in copyright protection technology. Platforms could also collaborate with copyright holders by jointly establishing copyright libraries or outsourcing copyright management to reduce the pressure of technical investment and ensure proper management and legal use of fast drama copyrights.

Ultimately, as consumers' and creators' awareness of copyright increases, platforms will face growing external pressure to consciously fulfill their review responsibilities. This external pressure comes not only from the strengthening of legal regulations but also from the market competition that influences platform brand image and user trust. By fulfilling their review responsibilities, platforms encourage creators to pay more attention to protecting intellectual property rights, while consumers' heightened copyright awareness fosters the healthy development of the fast drama industry. This positive cycle will attract more creators to choose the platform, enabling the platform to acquire more content, attract more consumers, and foster a virtuous cycle of creation, release, and consumption. This will ultimately bring increased profits to the platform and promote a multi-party win-win scenario for the industry, platforms, creators, and consumers.

Therefore, in the current technological environment, platforms should establish reasonable review standards, fulfill their prior review obligations, and initiate secondary review mechanisms when clear infringing behaviors are detected. If the platform fails to fulfill these responsibilities, it should bear the corresponding infringement liability to promote industry compliance and development.

4.3 Legal Regulation for Improving Copyright Protection Mechanisms for Fast drama

In light of the historical context behind the "safe harbor" principle, this principle, in the current era of highly developed algorithmic technologies, is evidently unable to comprehensively and effectively protect the rights of copyright holders [21]. With the rapid development of algorithmic technologies, platforms now possess stronger technical

capabilities, and their technological means and data analysis abilities are no longer limited to being a “neutral platform” for information. Therefore, they should assume the obligation and responsibility of prior copyright review for short video dramas. To address this, the relevant legal provisions can be improved in the following ways:

First, clarifying platforms' copyright review obligations. With the advancement of algorithmic recommendation technologies, platforms are shaping the dissemination and consumption of content through data analysis, user behavior prediction, and recommendation mechanisms. As a result, they are no longer merely information intermediaries, but are active guides to content consumption. To effectively protect creators’ rights, platforms must bear the responsibility of prior content review. This responsibility should not be limited to superficial checks of uploaded content, but should also include in-depth review and prevention of copyright violations. The law should explicitly require platforms to establish specialized copyright review departments, which are responsible for reviewing user-uploaded fast drama content and its associated materials. Platforms should implement clear legal requirements and industry standards in their copyright review processes, ensuring that they comply with copyright regulations before content is distributed, and strictly adhere to data protection and privacy laws.

Second, defining the platform’s “should know” standard and establishing a secondary review mechanism. The platform’s “should know” standard is no longer a simple judgment issue, especially when it comes to infringing fast dramas that may evade detection by review systems. When the content reaches a certain level of dissemination or interaction (such as views or shares), platforms must assess whether they have reached the “should know” standard and activate a secondary review mechanism. To facilitate this, platforms should regularly publish the criteria for triggering the secondary review mechanism and report these to industry regulators, thus enhancing public oversight and preventing platforms from using technological means to evade responsibility.

Third, strengthening platform responsibility, especially regarding the dissemination of infringing content. Due to platforms’ technological capabilities and content recommendation mechanisms, they play a decisive role in the spread of infringing content. Platforms accelerate the dissemination of infringing content through recommendations, trending topics, and push notifications, and their highly personalized recommendation systems can even exacerbate infringing activities. The law should clearly specify that platforms cannot rely solely on user reports or post hoc reviews. Instead, they must proactively identify and prevent the spread of infringing content. Platforms should bear

strict review and management responsibilities in every stage of content dissemination, including but not limited to recommendations, push notifications, trending topics, and content display in video feeds. If platforms fail to fulfill their review obligations or use technical means to circumvent review responsibilities, they should be held legally accountable.

Fourth, defining the timeframe for activating the secondary review mechanism. To ensure platforms promptly fulfill their review obligations, the law should specify the timeframe for activating the secondary review mechanism. For instance, within 24 hours after content reaches the specified level of dissemination or interaction, platforms should trigger the secondary review mechanism. If they fail to do so, they should bear corresponding legal responsibility. The legal consequences for failing to fulfill review obligations should include enhanced compensation, administrative fines, business restrictions, and other penalties.

5 Conclusion

Every fast drama, from creation and filming to market release, embodies the dedication and effort of its creators. Protecting the copyright of fast drama creators is not only about safeguarding their economic interests but also about encouraging innovation, promoting the prosperity of the cultural industry, and maintaining a well-regulated market order. Since 2007, YouTube has invested in developing the Content ID copyright system, with total investments exceeding $100 million. However, compared to its annual revenue of over $10 billion and the greater growth and revenue potential gained from protecting creators' copyright interests, this investment is undoubtedly necessary and worthwhile. This investment not only helps YouTube resolve long-standing copyright disputes but also fosters continuous content creation and supports the platform's healthy development. In today's era of rapidly advancing algorithmic technology, whether a platform can use AI algorithms for the pre-screening of fast drama copyrights is no longer a matter of technical limitation but rather a matter of technical choice. Under the "safe harbor" principle, the passive role of platforms is detrimental to resolving the increasingly complex and massive copyright disputes. Protecting intellectual property and the rights of copyright holders is also one of the core objectives of intellectual property law. Therefore, it is necessary to strengthen platform responsibility and establish platforms' pre-screening obligations for short drama copyrights, in order to

better protect creators' rights and promote the long-term development of intellectual property protection.

References

1. Yang, Y. : What Should Be Done about Copyright Protection and Development in the Fast drama Industry? Experts and Scholars Actively Discuss. China National Radio, November 29, 2024. https://www.sohu.com/a/831703619_362042? Last accessed 2025/1/23
2. Li, Z.: Short but Not Shallow, 'Free' but More Refined—Observing the New Developments in the Free Model of the Fast drama Market, Xinhua News, November 29, 2024. http://www.news.cn/fortune/20241129/dc0a8adc1018451d8c4f97841f60b1c3/c.html. Last accessed 2025/1/17
3. Luo, X.: Observing the 'Chaos' in Fast dramas: How Has the Industry Changed After the 'Strictest New Regulations for Short Dramas'?. Southern Metropolis Daily, July 26, 2024. https://finance.eastmoney.com/a/202407263141075368.html. Last accessed 2024/12/23
4. Wang, Q.: Viacom v. YouTube Case: When Will the 'Red Flag' Fly? - Also Commenting on the Impact of This Case on Video-Sharing Websites in China. vol. 4, pp. 4. China Copyright (2010). https://www.zhangqiaokeyan.com/academic-journal-cn_china-copyright_thesis/020125867501.html
5. Ma, Y., Zhang, R.: Theoretical Analysis of Network Service Providers' Duty of Proactive Review in Copyright Protection. vol. 21, pp.33-36. China Publishing (2018). doi:10.3969/j.issn.1002-4166.2018.21.009
6. Zhang, Q.: Dilemmas and Solutions in Identifying Content Infringement on Short Video Platforms. vol. 1, pp.44-46+48. Media (2022). https://qikan.cqvip.com/Qikan/Article/Detail?id=7106408253
7. Zhou, S.: Research on the Duty of Care of Short Video Platforms in the 'First Algorithm Recommendation Case.' vol. 4, 64-72. Journalists (2023). doi:10.16057/j.cnki.31-1171/g2.2023.04.003
8. Xiong, Q.: The Rules of Joint Infringement Between 'Algorithmic Push' and Network Service Providers. vol. 4, pp. 125-136. China Applied Law Journal (2020). https://qikan.cqvip.com/Qikan/Article/Detail?id=7102685233
9. See Hangzhou Internet Court Judgment (2021) Zhe 0192 Minchu No. 10493.
10. Yu, B.: On the Rationality of Network Intermediary Service Providers Bearing Review Obligations. vol. 1, pp. 169-175. Lanzhou Journal (2014). https://www.doc88.com/p-6681142489356.html

11. Luo, B., Song, S.: The Legal Nature of Algorithmic News Transmission Subjects: ICP or ISP—Also Discussing the Article 'Legal Perspective on Algorithmic News Recommendations.' vol. 6, pp. 77-86. Journalists (2019). doi:10.16057/j.cnki.31-1171/g2.2019.06.007
12. See Beijing Intellectual Property Court Judgment (2021) Jing 73 Minzhong No. 4293.
13. Zhang, M., Yang, H.: Review Obligations of Short Video Platforms in Short Video Infringement Cases. vol. 3, pp.88-90. Lanzhou Journal (2023). https://www.zhangqiaokeyan.com/academic-journal-cn_lanzhou-academic-journal_thesis/02012100981555.html
14. Li, Q., Wei, X.: The Rational Justification of the Limited Review System for Copyright on Short Video Sharing Platforms. vol. 3, pp. 65-68. Television Studies (2023). https://qikan.cqvip.com/Qikan/Article/Detail?id=00002H8G477G7JP0MJDO6JP16DR
15. Xu, J.: Research on the Duty of Care in Determining Copyright Infringement in Short Video Platforms from an Industrial Perspective. vol. 9, pp.31-40. Intellectual Property (2021). doi:3969/j.issn.1003-0476.2021.09.003
16. Cui, G.: On the Copyright Filtering Obligation of Network Service Providers. vol. 2, pp. 215-237. China Legal Studies (2017). doi:10.14111/j.cnki.zgfx.2017.02.011
17. Copyright Directive (2019),2019/790,2019 O.J(L130), art. 17 (4). https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=oj:JOL_2019_130_R_0004
18. Xu, J.: Fast dramas: The Winds Are Rising Again, TMTPost App, January 4, 2024. https://baijiahao.baidu.com/s?id=1786955727786955311&wfr=spider&for=pc
19. "Top 10 Short Drama Platforms in the New Influencer Power Rankings: Web Novel Platforms Incubating, MCN Shifting, Copyright Distribution Integrated," Sina Finance, September 17, 2024. https://baijiahao.baidu.com/s?id=1810457580209541132&wfr=spider&for=pc
20. Zhang, C., Ma, N.: A Study on the Proactive Review Obligation of Copyright for Internet Short Video Platforms. vol. 5, pp. 132-137. Modern Communication (China University of Communication Journal) (2021). doi:10.3969/j.issn.1007-8770.2021.05.023
21. Ma, Y., Zhang, R.: Theoretical Analysis of Network Service Providers' Duty of Proactive Review in Copyright Protection. vol. 21, pp. 33-36. China Publishing (2018). doi:10.3969/j.issn.1002-4166.2018.21.009

Data Science and Artificial Intelligence for Justice Delivery in India: Overview and Research Issues

Pavan Parvatam[1], P. Krishna Reddy[1], Gaurang Patil[1], K.V.K. Santhy[2], and M. Kumara Swamy[3]

International Institute of Information Technology, Hyderabad, Telangana, India[1]
NALSAR University of Law, Hyderabad, Telangana India[2]
CMR Engineering College, Hyderabad, Telangana, India[3]

Abstract. Data science and artificial intelligence (DSAI) based methods can process massive amounts of data to extract useful knowledge and can be employed to build decision support systems in various domains. Like other domains, the legal systems in several countries are being digitized and there is a scope to build DSAI-based frameworks to improve the performance of legal systems. In the literature, research efforts are being made to investigate DSAI-based methods to improve justice delivery. The Indian legal system is currently experiencing a major problem with a substantial backlog of cases. This paper provides an overview of DSAI-based efforts in the legal domain related to India. We also listed potential research issues to be explored. We hope the issues will encourage further research to improve justice delivery performance in India and other countries.

Keywords: AI and Law · Legal Data Analytics · IT for Law

1 Introduction

In today's world, a vast amount of data is generated due to the digitization of various systems. Over the last three decades, data science and artificial intelligence (DSAI) based concepts such as database systems, data warehouse, data cube, pattern mining, clustering, classification, regression, and machine learning/artificial intelligence [12] have been developed to organize and process different types of data for building decision support systems (DSSs) and search systems in various domains. Like other domains, legal systems in several countries are being digitized. Efforts are being made to improve the performance of legal systems by employing the latest advancements in DSAI.

Globally, around 5.1 billion individuals have been estimated to face unresolved issues related to justice [24]. In particular, the Indian judicial system is under tremendous pressure due to the huge number of pending cases. As of January 2024, more than 50 million civil and criminal cases were awaiting resolution in India [2,45].

Judges, lawyers, students/interns, investigators, and the common public are the key stakeholders and users of the legal system. Globally, efforts are being

made to extend developments in DSAI to improve justice delivery. The size and complex domain-specific formats are significant issues in developing DSAI methods to build DSSs to improve the performance of stakeholders. Like other domains, the judiciary system in India has transformed from manual to digitization. As a result, judiciary related data is being generated at various levels in digital formats. In India, research efforts are also being made to extend DSAI approaches to improve justice delivery. Several DSAI-based tools are being developed in the public and private domain [10].

In this paper, we first provide an overview of the Indian judicial system and explain the processing of a legal case. We will review the DSAI-based efforts in the legal domain related to India. We also discuss the DSAI-based efforts abroad briefly. Finally, we list the potential research issues to be explored.

Even though this paper is written by considering the Indian legal system, we hope that the contents of this paper will help the DSAI researchers, who do not have a legal background, understand the terminologies of a typical legal system with the corresponding stakeholders and the processing of a legal case. It will also help the researchers of DSAI to visualize the frameworks of stakeholders-specific DSAI-based DSS. We also hope that the research problems will encourage DSAI researchers to collaborate with legal practitioners to conduct further research to improve the performance of justice delivery in India and other countries.

The organization of the paper is as follows. In the next section, we will explain the Indian justice delivery system. In Section 3, we explain the steps involved in processing the case. In Section 4, we review the related research in India and abroad. In Section 5, we list the potential research problems. The last section provides a summary and conclusion.

2 Overview of the Indian Justice Delivery System

In this section, we first briefly explain the court system in India. Next, we explain the types of cases and stakeholders/users.

2.1 Details of Court System

The Indian judiciary system is a hierarchical structure that operates according to the constitution of India and the rule of law at various levels of courts. Fig. 1(a) shows the hierarchical relationship among the courts. India is geographically divided into states, with each state further subdivided into districts. The Supreme Court is the highest court, also called the apex court. Each state contains a High Court, and each district contains civil and criminal courts (sessions) where trials are held. The important legal terms used in this paper are described in Table 1. We now briefly explain the different types of courts in India.

- **Lower Court**: Lower courts such as judicial magistrate and junior civil judge courts serve as the first point of legal redress, handling civil cases like property disputes and family matters, as well as criminal offenses under the *Bharatiya Nyaya Sanhita* (BNS).

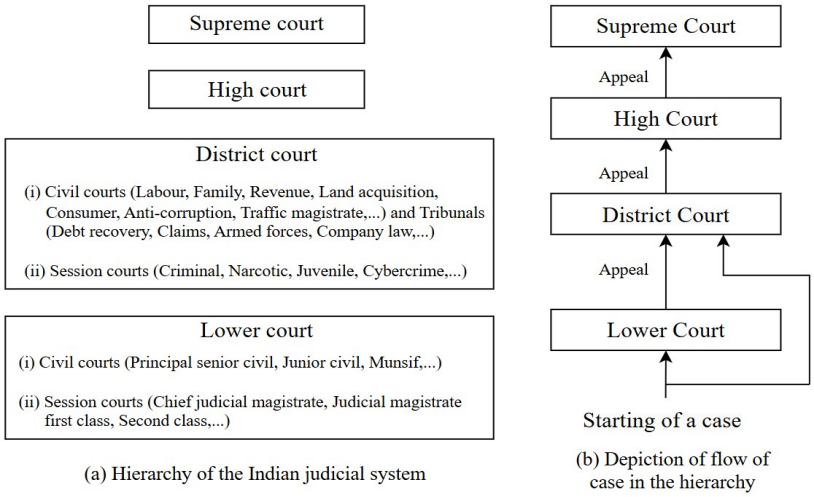

Fig. 1. The Structure and Case Flow in the Indian Judicial System

- **District Court**: The District and Sessions Courts handle various types of cases. Civil matters are addressed by District Courts, Sub-Courts, Principal Junior Civil Courts, and Munsif Courts. Criminal cases fall under Session Courts, led by the Sessions Judge under whom the Chief Judicial Magistrate, and first and second Class Judicial magistrates would be functioning. Special courts and tribunals, such as Family, Consumer, Labor, Tax Tribunals, and Administrative Tribunals, etc., focus on specific domains.
- **High Court**: A High Court is a judicial body that typically exercises jurisdiction over a state or a union territory. The appeals from district courts are handled by the High Court.
- **Supreme Court:** The Supreme Court is the apex court in India and typically hears appeals from the High courts across India.

2.2 Types of Cases

There are broadly two types of cases, as described below:

- **Civil Case**: A civil legal case is a legal dispute between two or more parties regarding a civil wrong. The disputes include issues relating to contracts, properties, family law, employment, consumer protection, etc.
- **Criminal Case**: A criminal case is a case that involves a person accused of committing a crime where one party to the crime is the State, and the ultimate goal is to punish the criminal or accused. The crimes include murder, rape, robbery, accidents, kidnapping, arson, other anti-social activities, etc.

Fig. 1(b) depicts the flow of a legal case. A plaintiff has to approach the lower court or district court in case of a dispute. The trial of the case happens at the

Table 1. Description of key legal terms

Term	Description
Affidavit	A statement written as an oath of truthfulness.
Appeal	A request for a court to review and overturn a decision made by a court or administrative body.
Arrest	A document issued by judicial authority to arrest someone over an offence.
Bail	Temporary release of an accused person with special conditions.
BNS	"Bharatiya Nyaya Sanhita" is the official criminal code in India, comprises of various chapters and sections related to Indian Law.
Charge-sheet	A document detailing the charges against a person accused of committing an offense by the Investigating officer.
Civil case	A legal case involving disputes between individuals or organizations over civil wrongs.
Criminal case	A legal case involving charges against a person accused of committing a criminal offense.
FIR	First Information Report (FIR) is a document created by police to record the details of a crime.
Hearing	A hearing is a formal proceeding in a court or other decision-making body where evidence and arguments are presented to decide a case. Hearings can be civil or criminal, and can take place before a judge, magistrate, or other decision-maker.
Lawsuit	A lawsuit is a civil legal action by one person or entity (the plaintiff) against another person or entity (the defendant), to be decided in a court.
Legal case	It is a dispute between two parties that is brought before a court to be resolved through a legal process
Order	A written directive issued by a court.
Summons	A legal document issued by a court requiring a person to appear in court.
Trial	A trial is a legal process where parties present evidence in a court to resolve a dispute. The trial is conducted by a judge or jury, who weigh the evidence and the law to reach a decision.

lower court or district court. Appeals from lower courts at the district level are dealt with by the higher courts at the district level. High courts only deal with appeals from the district courts. The Supreme Court only deals with appeals from the high courts.

2.3 Key Stakeholders

This section explains the stakeholders or users of a legal decision support system.

– **Lawyer**: We first explain the plaintiff and defendant and then explain about the lawyer. In a civil case, a plaintiff is a person or entity that files a lawsuit against another person or entity, the defendant. In a criminal case, the state or people prosecute the defendant for a crime against society. A defendant is a person or entity against whom a criminal or civil action is brought.

Lawyers represent their clients (plaintiff or defendant) in court and other legal forums and work to protect their rights and interests. Lawyers often help their clients resolve disputes through negotiation and settlement rather than going to court. A prosecutor is a lawyer appointed by the court to fight for the victim in a criminal case on behalf of the State (since the State is mandatorily one party to a criminal case).

- **Judge**: A judge is a person who presides over court proceedings, either alone or as a part of a panel of judges. India follows the adversarial system of justice delivery system where the judge hears all the witnesses and any other evidence presented by the lawyers of the case, assesses the credibility and arguments of the parties, and then issues a ruling in the case based on their interpretation of the law and their judgment. A judge is expected to conduct the trial impartially in an open Court.
- **Police (Investigating Officer)**: Police are primarily responsible for conducting investigations apart from other special investigating agencies such as the Narcotic Bureau, Food Inspectors, the Enforcement Directorate, etc. They gather evidence, examine witnesses, arrest suspects, and collect necessary evidence to assist the court in conducting trials.
- **Forensic experts**: Forensic experts analyze evidence associated with a case using scientific methods to aid investigations and court cases.
- **Law students and interns**: LLB (Bachelor of Legislative Law) students are often called law students or aspiring lawyers. The LLB is the first professional degree in law. Upon completion of LLB, graduates can pursue a legal career as lawyers, judges, legal advisors, and other legal professionals. A law intern is a law student or recent law graduate who participates in an internship program to gain practical experience. The skills they gain include researching, drafting documents, collecting evidence, and preparing for cases.
- **Law researchers**: Law researchers analyze legal documents, case laws, and other legal matters to help answer legal questions. They work in law firms, government agencies, non-profit organizations, and corporate legal departments. It is an essential skill for working on any case. It helps lawyers provide well-informed advice to clients and ensure sound decision-making.
- **Common Public**: The common public typically expects a law system to be accessible, fair, impartial, efficient, transparent, and timely in resolving disputes, meaning everyone should be able to access justice, receive without bias, have cases resolved promptly, and understand the reasoning behind legal decisions, all while being free from corruption or undue influence.

3 Processing of the Legal Case

In this section, we explain the typical flow of the case and then explain the important steps in the proceedings of the civil and criminal cases. Next, we present the details of dispatch latency in a civil and criminal case.

Fig. 2 illustrates how the arguments take place between the defendant and the appellant with the help of lawyers defending their statements in the court

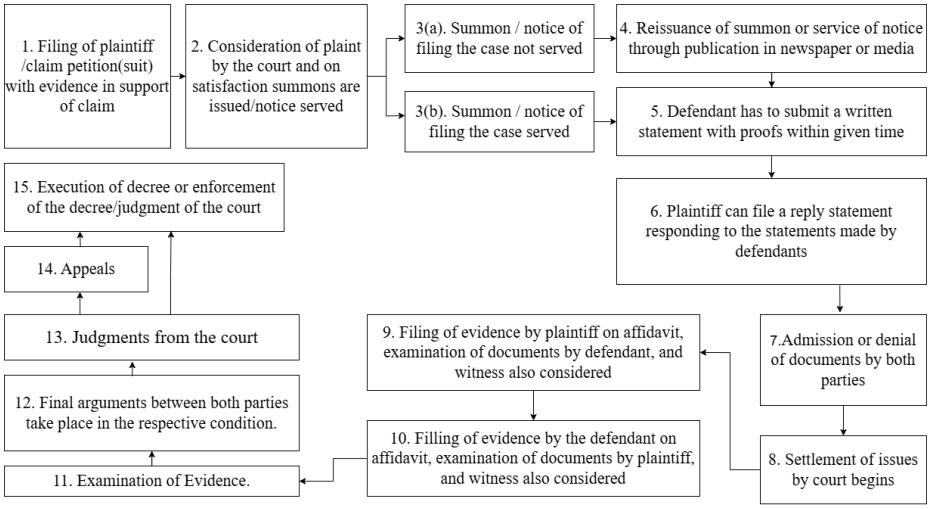

Fig. 2. Detailed steps of a legal case

with multiple arguments before final judgment in the presence of judge [1]. It starts with the plaintiff filing the lawsuit. The exchange of notices and replies occurs between the plaintiff and the defendant (Steps 1 to 7). Next, the court examines the evidence (Steps 8 to 11). Subsequently, the arguments in the court take place (Step 12), and the judgment is delivered (Step 13). If any party (plaintiff or defendant) appeals, the arguments happen at the higher court (Step 14). Otherwise, the court order/decree enforcement occurs (Step 15).

Fig. 3 depicts the case flow by identifying important phases. Both civil and criminal case proceedings are divided into two phases: the pre-trial phase and the trial phase. The pre-trial phase of a civil case (Fig. 3(a)) starts with the plaintiff's complaint to the court through a lawyer. Next, the court sends the notice to the defendant. Subsequently, the defendant, through a lawyer, responds to the notice. The pre-trial phase (Fig. 3(b)) of a criminal case starts with the complaint to the police by the plaintiff (or victim). Next, the police file an FIR and submit it to the court. The subsequent action takes place based on the court's direction.

For both civil and criminal cases, the trial phase (Fig. 3(c)) consists of several hearings. In each hearing, both lawyers (plaintiff, defendant, witnesses) participate in the arguments before the judge, and the judge may pass an interim order. The final judgment is delivered after completing the trial phase.

Based on the preceding explanation, the dispatch latency details of civil and criminal case proceedings are depicted in Fig 4. The total case time of civil and criminal case proceedings is divided into pre-trial and trial phases. The trial phase is divided into the discovery phase (exchange of documents, statements, and evidence) and hearings phase in civil case proceedings. In the case of a

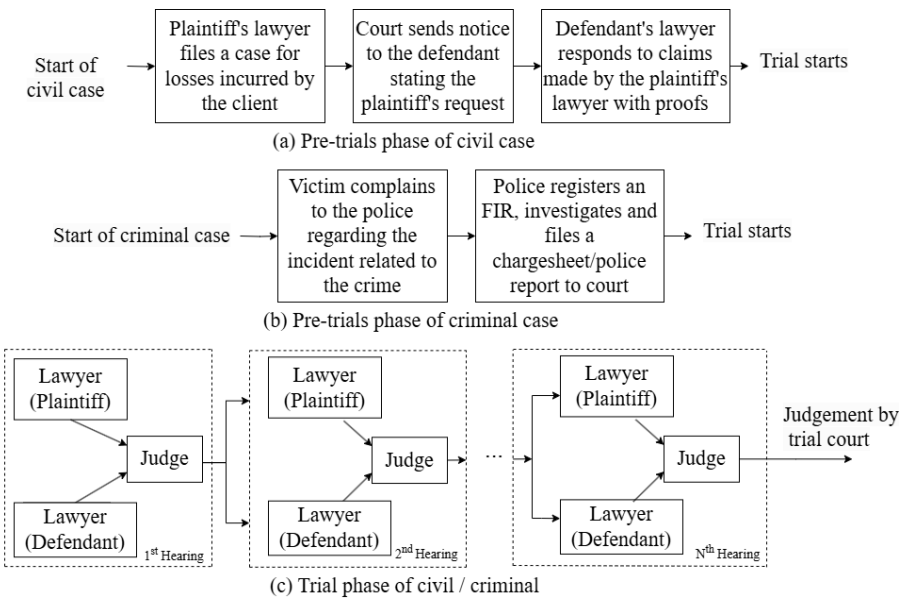

Fig. 3. Depiction of process at the trial court

criminal case, the pre-trial phase consists of arrest and case investigation. The trial phase consists of multiple hearings.

Discussion on delays: In case of a dispute, a plaintiff/victim approaches the lower court/district court, i.e., the justice system, to receive justice. An ideal justice system should deliver justice to the plaintiff at the earliest possible. However, especially in the Indian scenario, justice delivery is delayed due to several types of delays [20] for processing the case. Examples of delays include postal delays, investigation delays during the pretrial phase, and scheduling delays during the trial phase due to overburdened courts. In case of appeal, the delay is further prolonged.

4 Related Research

In this section, we first present the related DSAI research to improve Indian law. Next, we briefly summarize the DSAI research trends abroad. We provide concrete observations about DSAI research in India.

4.1 About DSAI-based law research in India

Some of the early works in the legal domain included, for example, tasks like summarization of legal documents [37,38,39] and finding similarity between legal documents [21]. A graphical model for legal document summarization was

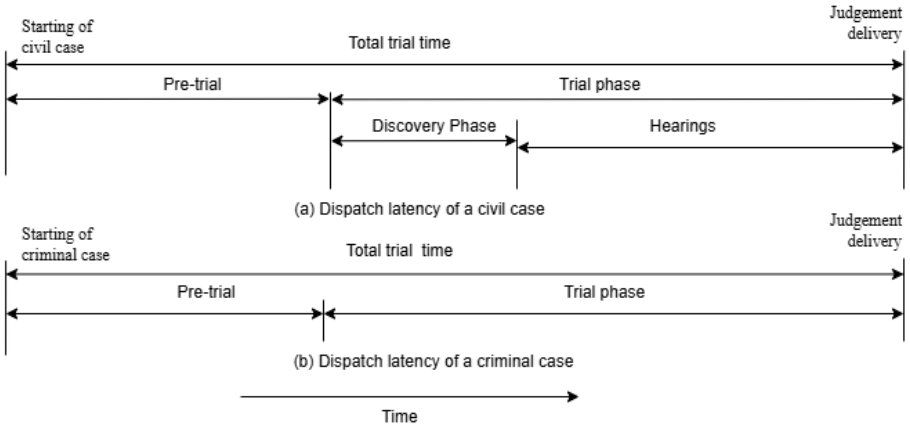

Fig. 4. Dispatch latency of a civil and criminal case: starting of the case to judgment delivery

proposed in [39]. In [37], authors leveraged conditional random fields to identify rhetorical roles for the summarization task. Authors in [21] showed that citation networks could be leveraged for estimating similarity between legal documents and found bibliographic coupling effective.

Over the years, with significant advancements in NLP, Information Retrieval and Deep Learning, the focus has broadened and now focuses on developing AI-based solutions for more challenging and sophisticated tasks. In recent years, the Indian research community has made efforts to apply advanced AI-based techniques in the domain of justice delivery and law and has also established some benchmarks for the Indian legal domain [13,17]. The authors in [13] have developed the IL-TUR benchmark for facilitating NLP research in the Indian legal domain. They have also set several LLM-based baselines for legal NLP research. We provide brief explanations for the tasks below.

- **Legal Information Retrieval:** One of the important problems in legal information retrieval is precedent retrieval. Legal practitioners are often required to cite precedents while formulating arguments in court and while drafting judgments. Manually retrieving relevant precedents is a very tedious task. Several approaches to automate the task have been proposed. Event-based representations have been proven effective for retrieval [14]. A number of possible document representations have also been experimented with [30,42]. Citation networks have been utilized extensively for finding similar cases [6,31]. Approaches leveraging both the citation network and text-based similarity have also proven effective [7,26,34].
- **Summarization of Legal Documents:** Summarizing legal documents is challenging due to their length and the intricate nature of legal cases. Summarizing legal documents aids legal practitioners. Efforts have been made to create benchmark datasets for extractive and abstractive summarization of

Indian court judgments [27,41]. It has been found that fine-tuning Legal-LED for abstractive summarization of Indian judgments is effective [13].

- **Named Entity Recognition (NER):** Legal NER involves the identification of legal entities in legal documents. Sample legal entities include the appellant, defendant, dates, statute, case number, etc. Developing models for NER is an important problem, and efforts have been made to create datasets and to develop transformer-based models for NER [16].
- **Rhetorical Role Labeling:** Rhetorical Role Labeling involves identifying rhetorical roles for each sentence in a given legal document. Examples of rhetorical roles include - facts, arguments, ruling by the court, the ratio of decision, etc. [8]. Identifying the rhetorical roles of sentences can help improve legal search [33] and summarization of judgments [37]. Authors in [8,9] developed deep learning models for identifying seven rhetorical roles. Authors in [28] focused on developing models for fine-grained rhetorical role labelling involving 13 class labels.
- **Judgment Prediction and Explanation:** This judgment prediction task typically involves a binary classification task where- given the facts and arguments associated with a case, the model predicts if the appeal from the petitioner is to be accepted or dismissed, along with generating an explanation for the same [29]. In [44], the authors evaluate the performance of various LLMs in the judgment prediction and explanation task and found that LLMs are good at generating explanations. In [19], the authors explored the applicability of GNNs for the prediction task.
- **Legal Statute Identification:** Legal statute identification involves identifying the statutes applicable to the case at hand. Earlier, the statutes for criminal law were described in the form of the IPC (Indian Penal Code), which has now been replaced by the *Bharatiya Nyaya Sanhita*. Such tools can help legal practitioners quickly identify relevant parts of the law that apply to a given case. In [35], a GNN-based approach was proposed for legal statute identification. Efforts have also been made to create datasets for explainable legal statute identification [44]. In [44], the authors find that LLMs are good at statute identification tasks while being moderately good at generating explanations for the same.
- **Bail Prediction:** Bail prediction is a task mainly designed for district courts in India and has been framed as a binary classification task, where, given a bail application, the model decides whether the bail should be granted or dismissed [18]. In [18], the authors release a dataset comprising legal documents written in Hindi and propose a Multi-Task Learning-based model for the bail prediction task. In [4], authors leveraged CNN for the same task.
- **Translation of legal documents:** Several Indians are more comfortable reading in regional languages, and there have been efforts to create datasets for the translation of judgments from English to the local languages [25]. In [25], authors released benchmark datasets for the translation of the legal text to Indic languages and carried out extensive experiments using various models (including LLMs) for the translation task.

Apart from the preceding tasks, there have been efforts to build a fine-tuned LLM for the Indian legal domain, called Aalap, with the primary focus being training the model for legal reasoning [43].

4.2 About DSAI and Law Research Abroad

International conferences and journals focusing on advancements at the intersection of AI and law include, for example, JURIX (International Conference on Legal Knowledge and Information Systems), ICAIL (International Conference on Artificial Intelligence and Law) and Journal of Artificial Intelligence and Law. Some of the trending research areas among researchers in foreign countries include legal reasoning and argumentation [23,32], knowledge representation [3,40] and applications of LLMs in the legal domain [5,36]. In particular, leveraging AI for legal reasoning has been a topic of interest for decades [11]. In [22], authors have surveyed applications of legal LLMs, elaborating on possible use cases, challenges, and future research directions. It can be observed that the applications of legal LLMs are a topic of great interest among researchers.

4.3 Observation

Overall, it can be observed that most of the existing literature in the Indian legal domain focuses on creating benchmark datasets and developing and improving models for various NLP and IR-related tasks. However, certain issues arising at different stages of the case flow, which contribute to delays in the delivery of justice, have not received sufficient attention from the research community. A lack of comprehensive research on stakeholder-specific problems can also be observed.

Researchers from foreign countries have focused on problems such as legal reasoning, argumentation, and the application of large language models (LLMs) in the legal domain, owing to the differing nature of challenges faced by their respective countries. More stakeholder-specific research leveraging DSAI is needed to enhance the performance of the Indian judicial system.

5 List of Research Issues

As observed in the preceding section, there are some past and ongoing efforts (not many) to extend DSAI approaches to improve the performance of the Indian judicial system. Earlier, the research mainly leveraged data sets comprising legal documents and focused on information retrieval, precedent retrieval, and summarization. Recently, there have been efforts to leverage LLMs for various tasks. With this background, we provide the research issues identified in the workshop[1] by brainstorming between DSAI researchers and experts from the legal domain.

[1] A workshop entitled "Data Science for Justice Delivery in India (DSJDI2022)" was held on December 10, 2022, at IIIT Hyderabad in conjunction with 10th International Conference on Big Data Analytics, 2022 (BDA2022)

1. **Autonomous judicial system:** The pre-trial and trial phase of a case is often prolonged due to numerous reasons. Although some repetitive tasks can be easily automated, complex tasks require extensive research. The goal of building an autonomous judicial system is very challenging. Creating a goal for building an autonomous judicial system will result in many valuable tools that may not replace the existing judicial system but assist in many ways that can help deal with certain tasks autonomously.
2. **Judicial precedents for legal reference:** Legal practitioners often need to cite precedents to support their arguments. Manually retrieving relevant cases is time-consuming and inefficient. Building precedent retrieval systems to help reduce the workload of lawyers and judges is currently an active research area.
3. **Language translation:** India is a multilingual country, and several courts operate in regional languages. Translation between different languages is necessary to make the justice system accessible to stakeholders across the country. Advancements in the domain of NLP have made the problem tractable. The "SUVAS" system adopted by the Supreme Court of India involves the translation of judgments into some Indian regional languages[15]. This can be improvised and scaled to accommodate other languages and can be made available at other courts in the judicial hierarchy.
4. **Analysis of judgments:** Court judgments need to be analyzed at the macro level, which may, for example, include the study of crimes committed and how they are growing or the study of cases related to divorces, suicides, atrocities, etc. Such an analysis not only gives us a way to evaluate and understand our judicial system but also helps us understand society and its problems. Such a system may be relied upon to develop policies based on the emerging trends in society that strive to improve thinking and provide adaptability for the common public in India.
5. **Digitization of trial court data:** Several trial courts in the country may not be fully digitized. Argument and judgment data can potentially be a valuable resource for researchers and can provide valuable insights. A major effort must be made to digitize the historical argument data and the judgment data of all trial courts in India while preserving privacy. Such data can also be utilized by other courts.
6. **Summarization of documents:** Trials involve a lot of paperwork, and stakeholders are often required to read very long documents. Reading such documents and extracting the most relevant information from the document is often cumbersome. Summarization of legal documents can help legal practitioners understand the document in brief and its automation can decrease their workload.
7. **Legal document drafting:** Legal practitioners are often required to draft case documents, agreements, contracts, etc. Leveraging DSAI-enabled systems to assist in drafting the document can reduce the workload of lawyers, judges, and other stakeholders. Additionally, such tools can be used for verification and to find possible flaws in documents. Possible flaws may include

spelling or grammatical errors and logical errors that can be difficult to identify.

8. **Simplification of legal documents:** Understanding and interpreting legal documents can be very challenging for the common public. Simplification of legal documents is very crucial for the law to be more accessible. NLP techniques for simplifying legal documents need to be explored.
9. **The platform of case management and tracking:** A case management platform should be built to connect all of the stakeholders in the justice delivery system; such systems can increase efficiency as the information can be exchanged faster and in a secure manner. Such a system can also help maintain the chain of events and evidence, thus providing a better case overview and fast-track justice delivery.
10. **Organizing court proceedings and scheduling hearings:** In order to make legal case management smoother, the AI system can be equipped with details about court hierarchies and judge benches. This would help to efficiently schedule hearings, reduce delays, and minimize errors. The system can also help build trust in the legal process by verifying the authenticity of evidence and cross-checking documents and transcripts in lower courts and other official records.
11. **Virtual hearing:** Virtual hearings are present in the Supreme Court and High Courts but not in the lower courts. Virtual hearings in lower courts can help to fast-track justice delivery. Implementing virtual hearings in lower courts in India can significantly reduce case backlogs by expediting routine matters, bail hearings, and preliminary arguments without unnecessary delays. This enhances access to justice, especially for individuals in remote areas, by minimizing travel costs and logistical challenges faced by litigants, lawyers, and witnesses
12. **Lack of standardization:** Judicial documents often lack standardization that can help fast-track legal procedures. Implementing standardized templates for orders, judgments, and filings can streamline case processing, reduce clerical errors, and improve inter-court coordination. Judicial documents should be required to follow a standardized structure to ensure consistency across states and languages.
13. **Practical training for law interns and students:** Innovative DSAI-based tutoring systems can be leveraged to impart practical skills to law interns and students to produce more competent legal practitioners. It can provide hands-on legal training to Indian law students, bridging the gap between theory and practice. This can enhance courtroom readiness and improve the efficiency of legal research.
14. **Human resource management:** Lack of coordination among stakeholders is a significant issue in the judicial system. Building databases to keep all stakeholders updated with the necessary documents and information can fast-track the delivery of justice. AI can also play a crucial role in resource management.
15. **Disparity in judgments:** Decisions taken by courts in similar cases may have disparities. Among the many reasons for the stated disparity, gender

bias is predominant. Inconsistent court rulings in similar cases can create uncertainty in the legal system. Gender bias, particularly in cases related to marital disputes, workplace harassment, and inheritance rights, often leads to varying interpretations.

16. **Judicial values and societal morality:** While AI often provides accurate and comprehensive analytics, it is crucial to understand the limitations that must be imposed on its usage to ensure that judicial values, constitutional principles, and societal morality are not compromised.
17. **Corruption-free, robust systems:** Systems designed for the judiciary need to be resistant to external influences, such as corruption and political pressure. Cryptography offers various methods to make systems more secure, trustworthy, and reliable. A transparent and digitally secure legal system that can reduce bribery, case manipulation, and delays and ensure trustworthy and corruption-free justice delivery should be implemented.

6 Conclusion

The Indian judicial system is overburdened with a huge number of pending cases. There is an opportunity to exploit recent developments in DSAI to improve the situation. In this paper, we provided potential research issues after an overview of the DSAI-based efforts in the legal domain related to India. Notably, more research is required to investigate decision support systems by considering users' requirements, such as judges, lawyers, law interns, students, and investigators. Most importantly, an effort has to be made to digitize the historical argument data along with the judgment data of all trial courts in India while preserving privacy. Such an effort will enable the building of tools to provide stakeholder-specific services to improve the Indian justice system.

Acknowledgments: We acknowledge the support of iHub Anubhuti-IIITD Foundation set-up under the NM-ICPS scheme of the Department of Science and Technology, India. We also thank Sri M. Radha Krishna Chauhan (Honorable judge), U. Narendra Babu, and Bhoomeendra Singh Sisodiya for their help in conducting the brainstorming workshop.

References

1. The code of civil procedure, 1908, `https://www.indiacode.nic.in/bitstream/123456789/2191/1/A1908-05.pdf`, accessed: 2025-01-31
2. National Judicial Data Grid (NJDG), `https://njdg.ecourts.gov.in/`, accessed: 2024-01-13
3. Atkinson, K., Bench-Capon, T.: Angelic ii: An improved methodology for representing legal domain knowledge. ICAIL '23, ACM (2023)
4. Barman, A., Roy, D., Paul, D., et al.: Convolutional neural networks can achieve binary bail judgement classification. ICON 2023, NLPAI (2023)
5. Belfathi, A., Hernandez, N., Monceaux, L.: Harnessing gpt-3.5-turbo for rhetorical role prediction in legal cases. In: JURIX. IOS Press (2023)

6. Bhattacharya, P., Ghosh, K., Pal, A., et al.: Hier-spcnet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. SIGIR, ACM (2020)
7. Bhattacharya, P., Ghosh, K., Pal, A., et al.: Legal case document similarity: You need both network and text. Information Processing & Management **59**(6) (2022)
8. Bhattacharya, P., Paul, S., Ghosh, K., et al.: Identification of rhetorical roles of sentences in indian legal judgments. In: JURIX. IOS Press (2019)
9. Bhattacharya, P., Paul, S., Ghosh, K., et al.: Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. Artificial Intelligence and Law (2023)
10. ContractSafe: Legal ai tools: Benefits, use cases, and examples (2025), https://www.contractsafe.com/blog/legal-ai-tools
11. Gardner, A.v.d.L.: An artificial intelligence approach to legal reasoning. MIT press (1987)
12. Han, J., Pei, J., Tong, H.: Data mining: concepts and techniques. Morgan kaufmann (2022)
13. Joshi, A., Paul, S., Sharma, A., et al.: IL-TUR: Benchmark for Indian legal text understanding and reasoning. In: ACL (2024)
14. Joshi, A., Sharma, A., Tanikella, S.K., et al.: U-CREAT: Unsupervised case retrieval using events extrAcTion. ACL (2023)
15. Judiciary, D.: Supreme court vidhi anuvaad software (suvas) (2025), https://www.drishtijudiciary.com/current-affairs/supreme-court-vidhik-anuvaad-software-suvas
16. Kalamkar, P., Agarwal, A., Tiwari, A., et al.: Named entity recognition in Indian court judgments. NLLP Workshop, ACL (2022)
17. Kalamkar, P., Venugopalan, J., Raghavan, V.: Benchmarks for indian legal nlp: a survey. In: JSAI International symposium on artificial intelligence. Springer (2021)
18. Kapoor, A., Dhawan, M., Goel, A., et al.: HLDC: Hindi legal documents corpus. ACL (2022)
19. Khatri, M., Yusuf, M., Kumar, Y., et al.: Exploring graph neural networks for indian legal judgment prediction. preprint arXiv:2310.12800 (2023)
20. Krishnan, J.K., Raj Kumar, C.: Delay in process, denial of justice: the jurisprudence and empirics of speedy trials in comparative perspective. Georgetown Journal of International Law **42**(3) (2011)
21. Kumar, S., Reddy, P.K., Reddy, V.B., et al.: Similarity analysis of legal judgments. In: Proc. of Fourth ACM Bangalore conference (2011)
22. Lai, J., Gan, W., Wu, J., et al.: Large language models in law: A survey. AI Open **5** (2024)
23. Liepiņa, R., Wyner, A., Sartor, G., et al.: Argumentation schemes for legal presumption of causality. ICAIL '23, ACM (2023)
24. Long, S.C., Ponce, A.: Measuring the justice gap: A people-centered assessment of unmet justice needs around the world. Tech. rep., World Justice Project (2019)
25. Mahapatra, S., Datta, D., Soni, S., et al.: Milpac: A novel benchmark for evaluating translation of legal text to indian languages. ACM TALLIP (2024)
26. Makawana, M., Mehta, R.G.: A novel network-based paragraph filtering technique for legal document similarity analysis. Artificial Intelligence and Law (2023)
27. Malik, M., Zhao, Z., Fonseca, M., et al.: Civilsum: A dataset for abstractive summarization of indian court decisions. SIGIR '24, ACM (2024)
28. Malik, V., Sanjay, R., Guha, S.K., et al.: Semantic segmentation of legal documents via rhetorical roles. NLLP Workshop 2022, ACL (2022)

29. Malik, V., Sanjay, R., Nigam, S.K., et al.: ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. IJCNLP, ACL (2021)
30. Mandal, A., Ghosh, K., Ghosh, S., et al.: Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law **29** (2021)
31. Minocha, A., Singh, N., Srivastava, A.: Finding relevant indian judgments using dispersion of citation network. WWW '15 Companion, ACM (2015)
32. Mumford, J., Atkinson, K., Bench-Capon, T.: Combining a legal knowledge model with machine learning for reasoning with legal cases. ICAIL '23, ACM (2023)
33. Nejadgholi, I., Bougueng, R., Witherspoon, S.: A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In: JURIX. IOS Press (2017)
34. Patil, G., Sisodiya, B.S., Reddy, P.K., et al.: Citation anchor text for improving precedent retrieval: An experimental study on indian legal documents. In: Legal Knowledge and Information Systems. IOS Press (2024)
35. Paul, S., Goyal, P., Ghosh, S.: Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents. AAAI (2022)
36. Ribary, M., Krause, P., Orban, M., et al.: Prompt engineering and provision of context in domain specific use of gpt. In: JURIX. IOS Press (2023)
37. Saravanan, M., Ravindran, B., Raman, S.: Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In: IJCNLP (2008)
38. Saravanan, M., Ravindran, B.: Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artificial Intelligence and Law **18** (2010)
39. Saravanan, M., Ravindran, B., Raman, S.: Improving legal document summarization using graphical models. Frontiers in Artificial Intelligence and Applications (2006)
40. Servantez, S., Lipka, N., Siu, A., et al.: Computable contracts by extracting obligation logic graphs. ICAIL '23, ACM (2023)
41. Shukla, A., Bhattacharya, P., Poddar, S., et al.: Legal case document summarization: Extractive and abstractive methods and their evaluation. In: AACL-IJCNLP (2022)
42. Sisodiya, B.S., Unnam, N.B., Reddy, P.K., et al.: Analysing the resourcefulness of the paragraph for precedence retrieval. ICAIL '23, ACM (2023)
43. Tiwari, A., Kalamkar, P., Banerjee, A., et al.: Aalap: Ai assistant for legal & paralegal functions in india. arXiv preprint arXiv:2402.01758 (2024)
44. Vats, S., Zope, A., De, S., et al.: LLMs – the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on Indian court cases. EMNLP 2023, ACL (2023)
45. Yasir, S.: 'a lifelong nightmare': Seeking justice in india's overwhelmed courts. New York Times (January 13 2024), `https://www.nytimes.com/2024/01/13/world/asia/india-judicial-backlog.html`, accessed: 2025-01-30

Enhancing Document Retrieval in Large Corpora: A Keyphrase and Reference-Based Approach

Zoltán Szoplák[0000−0003−1823−0536], Dávid Varga[0000−0002−3176−8106], Peter Gurský[0000−0002−4744−7390], Šimon Horvát[0000−0002−3191−8469], and Stanislav Krajči[0000−0001−5612−3534]

Institute of Computer Science, Pavol Jozef Šafárik University in Košice, Slovakia

Abstract. The task of document retrieval, especially in large corpora, remains a challenge despite the recent advances in NLP and LLM technology. The context windows for such models are often not sufficient enough to fit a large number of documents inside them. To handle such issues, many NLP tasks, such as document retrieval for RAGs, perform a two-step process of filtering out the most relevant documents and then using those as input for the LLM to rank them. However, the larger the corpus, the greater the need for a robust document retrieval method. We propose a solution that augments the standard embedding similarity retrieval method with a score based on matching extracted keyphrases and a further extension using references in legal documents. The former reached a mean average precision of 80.25% while the latter improved on the former's result even further, achieving a mAP score of 92.85%.

Keywords: document retrieval · keyphrases · legal references · information retrieval

1 Introduction

Document retrieval, especially in large text corpora, has been a significant problem in machine learning and natural language processing, especially in the recently popular field of Retrieval Augmented Generation (RAG). Despite the recent advances in the field of NLP and LLMs, retrieval tasks remain an issue due to the limited context window sizes of most LLM models. And doing it iteratively through an entire corpus is currently too costly to be feasible. Therefore, document retrieval systems usually operate in two stages. The first consists of selecting the most relevant set of documents based on embedding similarity which can then fit inside the context window of an LLM that would perform a reranking step. However, such approaches pose their own challenges. In the case of very large datasets, the embedding-based search may return many matches, making the reranking part of such algorithms computationally expensive. In the same vein, limiting the number of retrieved documents may cause relevant documents to be missed from the reranking step of the process. The size of the text can also cause major problems regarding such approaches. In extensive texts without a proper chunking algorithm, the uniqueness of a given text within a corpus can

easily be lost, especially when it comes to texts with a semi-formulaic structure, such as legal texts and judicial decisions. And while it is entirely possible to split the given document into chunks to avoid this problem, however chunking algorithms have their own challenges.

2 Related Work

Legal document retrieval is a specialized area of information retrieval (IR) that has evolved through advancements in natural language processing (NLP) and machine learning (ML). Key challenges include handling the complexity and length of legal texts, where paragraph-level retrieval [6] and embedding-based similarity models [2] improve access to relevant case law. Deep learning approaches, such as neural classifiers for legal opinions [7] and BERT-based case retrieval [1], further enhance retrieval accuracy by capturing nuanced relationships in legal texts.

Relevance feedback mechanisms play a critical role in refining retrieval results [9]. Pipitone and Alami [4] introduce LegalBench-RAG, a benchmark designed to assess retrieval performance in Retrieval-Augmented Generation (RAG) systems for legal texts, demonstrating the importance of precise document mapping and chunking strategies. Similarly, Wiratunga et al. [10] propose CBR-RAG, which incorporates Case-Based Reasoning (CBR) into RAG to improve retrieval by leveraging case indexing and similarity matching, reducing hallucinations in legal LLMs.

These advancements highlight the growing importance of integrating NLP, deep learning, and retrieval-specific enhancements to improve legal information systems. By combining effective chunking, domain-specific embeddings, and retrieval frameworks, modern systems achieve greater accuracy in retrieving relevant legal texts for case law analysis and question answering.

3 Methods

3.1 Overview

Before we dive into the details of our methods for retrieving legal documents, let us present the core parts of the algorithm. We can divide them into preprocessing and retrieval parts. Although some steps are marked as optional, they are quite beneficial for a quality result as will be discussed in section 4.2. Formal labels used in the outputs of the steps are defined in the following sections.

1. Preprocessing
   (a) Semantical chunking of all judicial decisions - output: list of chunks for each document.
   (b) Creating chunk embeddings via language model - output: vector embedding for each chunk $v(c_i)$
   (c) Extraction of keyphrases from chunks - output: top t keyphrases for each chunk - $KPP(c_i)$

i. Computing modified TF-IDF score for each keyphrase.
ii. Using Self-Attention MAPS (AttentionSeek) to create SAM score for each keyphrase.
iii. Combining the two metrics, extracting top t keyphrases from a chunk.

(d) Extraction of references from document chunks (optional) - output: list of referenced legal regulations and judicial decisions of each chunk.
(e) Computing IDF values of references - output: IDF value for each legal regulation and chunk (each chunk defers to the IDF value of the document it belongs to).
(f) Extraction of keyphrases from references (optional) - output: list of keyphrases for each legal regulation and each chunk of judicial decision.

2. Retrieval
(a) Creating embedding of query - output: embedding vector of query - $v(q)$
(b) Determining semantic similarity between query and chunk embeddings - output: cosine similarity between query and chunk - $cos(v(q), v(c_i))$
(c) Extracting keyphrases from the query - output: list of keyphrases from the query - $KPP(q)$
(d) Finding matching keyphrases between query keyphrases and chunk keyphrases - output: intersection between $KPP(q)$ and $KPP(c_i)$
(e) Combining the matching keyphrases with the cosine similarity and to get a relevance score of each document of judicial decision - output: $score(d_i)$
(f) Enhancing relevance score of documents by using cosine similarity and shared keyphrases of references as well. (optional) - output: final enhanced score of each document.

3.2 Basic labels

For our purposes, let's define the task of document retrieval as returning a list of documents $R = r_1, ..., r_m$ from a predefined corpus $D == d_1, ..., d_n$ where for $1 \leq k \leq m \leq n$, $r_k \in D$ with the most significant relevance to a given query text q as well as provide a ranking for said documents where for a given $1 \leq a, b \leq m \leq n$ if $similarity(q, r_a) \leq similarity(q, r_b)$ then $score(r_a) \leq score(r_b)$.

The most common method of achieving this lies in using some kind of language model to create vector embeddings $v(t)$ where t is a natural language text we expect our language model to be able to encode. We have opted to use the *kinit/slovakbert-sts-stsb* model, described in [3]. Using said language model, we are able to encode both the query q and the documents in D so that $score(d_i) = \cos(v(q), v(d_i))$ where $1 \leq i \leq n$, $d_i \in D$ and $\cos(v(q), v(d_i))$ denotes the cosine similarity between the embedding of the given document and the query.

3.3 Reference and keyphrase extraction

To improve upon existing document retrieval methods, we work within the core text and use various metadata to get more precise and specific results. Such

metadata includes references to law articles and previous court decisions, extracted using rule-based methods specific to Slovak Legal texts described in our earlier work [8].

These references can be used to perform our keyphrase extraction methods. However, we will present an approach to keyphrase extraction that doesn't require references yet still offers improvement over reranking.

Our keyphrase extraction method is inspired by the one first described in [5].

3.4 Extraction of potential keyphrases from the source document and its individual chunks

First, we establish a set of candidate keyphrases we will work with. In our case, that would be a set of phrases collected from different dictionaries into a vocabulary that will be consistent across the whole database and the operations we will perform over it. Let's denote this set of phrases as $P{=}p_1, ..., p_l$. This set of keyphrases has been manually approved by lawyers as the collection of terms and categories they wish to search for. Alternatively, one could generate candidate keyphrases based on POS tag pattern, other semantic and statistical metrics or simply take all explicit n-grams to a given degree from the target text.

Most keyphrase extraction methods, including ours, rely on the presence of a given term in the text document and its frequency. Such approaches have the downside of only being able to evaluate the phrases that directly appear in the document. To this end, we propose embedding similarity between individual phrases, and those phrases that pass a certain threshold of cosine similarity will be considered adjacent terms.

Let adj_i denote the set of phrases $\{p_j \mid p_j \cos(v(p_i), v(p_j)) \geq T$ where p_i and $p_j \in P$ and T is the threshold value.

When calculating the presence of a term, we use that specific term and all of its adjacent terms, weighted by the cosine similarity to the given term.

Let PP^t denote the present phrases within the text t. Then, for every i in P let

$$PP_i^t = \begin{cases} 1, & \text{if } p_i \in t \\ \cos(v(p_i), v(p_j)), & \text{if } p_i \notin t, p_j \in t \text{ and } p_j \in adj_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Then, we calculate TF-IDF on every keyphrase from a given set of keyphrases, using a maximum document frequency of 50% to exclude domain-specific stopwords and use sublinear frequency, which replaces the standard TF for $1 + \log(TF)$ to prevent the metric from overscoring with high IDF that appear far too many times in the document. We calculate the sublinear TF and IDF scores using the adjacent phrases.

Thus, TF-IDF$_{\text{mod}}$ will be caluclated as:

$$\text{TF-IDF}_{\text{mod}}(p_i, t) = 1 + \log\left(\text{TF-IDF}(p_i, t) + \sum_{p_j \in adj_i} PP_i^t \cdot \text{TF-IDF}(p_j, t)\right) \quad (2)$$

However, statistical methods like TF-IDF even with calculating frequencies of synonyms and related terms based on embedding similarity often fail to capture semantic information. For that reason, we decided to combine it with a keyphrase extraction method based on semantic features. For this purpose, we have chosen, AttentionSeek, a scoring method based on the Self-Attention Masks generating the given embedding, described in [11]. According to this method, we use Huggingface to extract the Self-Attention Masks (SAM) of our transformer model where A^{lh} will represent the attention scores for layer l and head h, where A_i^{lh} represents the attention score for layer l, head h and phrase $p_i \in P$ We create a so-called hypothesis vector H where for a given text t, $H = PP^t$ where $H_i = PP_i^t$. The relevance of a given attention vector is then calculated as the matrix multiplication for the A^{lh} of a given SAM and H vectors.

$$S^{lh} = A^{lh} \cdot H \quad (3)$$

We can calculate the relevance of each SAM, represented by R^{lh} as the average relevance of its attention vectors.

$$R^{lh} = \frac{1}{n} \sum_{i=0}^{n} S_i^{lh} \quad (4)$$

The final attention vector B^t for text t is computed as a weighted average across all SAMs' attention vectors.

$$B^t = \sum_{\forall l,h} \sum_{\forall i} A_i^{lh} * S_i^{lh} * R^{lh} \quad (5)$$

After we have calculated both of these keyphrase extraction metrics, we can create the combined measurement, the keyphrase potential KPP_i^t for a single keyphrase $p_i \in P$ in a given text t which consists of multiplying its modified TF-IDF score with the attention vector.

$$KPP_i^t = \text{TF-IDF}_{\text{mod}}(p_i, t) * B^t(p_i) \quad (6)$$

3.5 Using keyphrases for document retrieval

Let $D = d_1, .., d_n$ be the corpus of all judicial decisions. In the previous subsection, we described a way of extracting and ranking potential keyphrases from a given document. However, the texts of judicial decisions are usually long and must be split into chunks. There are several models of document chunking to consider. We found that fixed-size chunking was relatively inefficient regarding

judicial decisions. Due to the varying case complexities, the length of each document segment and semantic whole varies greatly. When it came to semantic chunking, we've had much greater success, particularly when establishing breakpoints through standard deviation, as portions of legal documents are, by their nature, more or less structured than others. Therefore, it yields greater success if we don't just consider semantic dissimilarity to the previous chunk but rather a normal change in the measure of dissimilarity within the said chunk.

Let $C(d_i) = c_{i,1}, ..., c_{i,j_i}$ be a set of all chunks for document $d_i \in D$ where $\forall_{1 \leq a \leq n} \forall_{1 \leq b \leq j_a}$

The relevant keyphrases for a given chunk can be extracted using the keyphrase extraction algorithm described in the subsection above.

$$KPP_i^{c_{a,b}} = \text{TF-IDF}_{\text{mod}}(p_i, c_{a,b}) * B^{c_{a,b}}(p_i) \tag{7}$$

Similarly, our query q may be a phrase, a collection of phrases, or a reference legal text; therefore, we can extract the keyphrases for the legal text the same way.

$$KPP_i^{q} = \text{TF-IDF}_{\text{mod}}(p_i, q) * B^{q}(p_i) \tag{8}$$

We can then select the top t keyphrases with the highest KPP scores.

Now that we have extracted the keyphrases, we need to note that not every chunk of text contributes equally to the final result. Therefore, we create a vector embedding of each chunk $v(d_{a,b})$ and calculate the cosine similarity between it and the vector embedding of our query $v(q)$. We multiply this similarity score with the number of shared keyphrases divided by the total number of keyphrases q and $d_{a,b}$ share.

$$score(d_i) = \frac{1}{|C(d_i)|} \sum_{c_j \in C(d_i)} \left(\cos\left(v(c_j), v(q)\right) \times \text{match}(c_j, q)\right) \tag{9}$$

where

$$\text{match}(c_j, q) = \frac{1 + 2 \times |top_t(KPP^{c_j} \cap top_t(KPP^q)|}{1 + |top_t(KPP^{c_j})| + |top_t(KPP^q)|} \tag{10}$$

3.6 Enhancing keyphrase augmented document retrieval using extracted references

Now that we know how to extract the keyphrases from a singular document, we can do the same for references. Let $\mathsf{L} = l_1, \ldots, l_d$ be the collection of all citeable legal regulations. Then let $\mathsf{L}_{\mathsf{ref}}(\mathsf{i}) = (l_{i1}, ...l_{in_i}) \subseteq \mathsf{L}$ the set of all legal regulations referenced by decision chunk $c_i \in D$.

Since these are short and to-the-point texts, they do not require any chunking, and their keyphrases can be extracted directly using the KPP^t formula.

However, not all references contribute equally to the meaning of the decision. Therefore, we calculate the IDF scores across all legal regulation references, as

those cited more often tend to represent standardized formal procedures more than something uniquely related to the document's content. Therefore, we calculate the IDF score for each legal regulation through the whole corpus.

We also consider how relevant a given legal regulation is to the chunk of text by examining the cosine similarity of its text to the chunk it was cited from.

$$KPP_k^{l_{ij_i}} = \cos(v(c_i), v(l_{ij_i})) \times IDF(l_{ij_i}) \times \text{TF-IDF}_{\text{mod}}(p_k, l_{ij_i}) \cdot B^{l_{ij_i}}(p_k) \quad (11)$$

As for court decision references, we can use the same formula as described, including multiplication by the IDF score and cosine similarity with the chunk it was cited from, except this step now has to be done for every individual chunk and averaged

Let the and $\mathsf{D_{ref}(i)} = (\mathsf{d_{i1}}, ...\mathsf{d_{in_i}} \subseteq \mathsf{D}$ the set of all court decisions referenced by decision chunk $c_i \in \mathsf{C(d_i)}$. Let $C(d_{ij_i}) = c_{ij_i,1}, ..., c_{ij_i,n_ij_i}$ be the set of all chunks from document $d_{ij_i}, i_1 \leq ij_i \leq in_i$.

Then the keyphrase score of d_{ij_i} be calculated as

$$KPPref_m^{d_{ij_i}} = \frac{1}{|C(d_{ij_i})|} \sum_{c_{ij_i,k_{ij_i}} \in d_{ij_i}} \cos\left(v(c_i), v(c_{ij_i,k_{ij_i}})\right) \cdot KPP_m^{d_{ij_i}} \quad (12)$$

where $KPP_m^{d_{ij_i}}$ is calculated using equation 7.

Now that we have the keyphrase scores for the references calculated, we can perform an element-wise addition to the keyphrase scores extracted from the core document using equation 7. For a given keyphrase with the index of m where $p_m \in P$, we calculate the score the following way:

$$KPPfinal_m^{c_i} = KPP_m^{c_i} + \sum_{l_j \in L_{ref}(i)} KPP_m^{l_j} + \sum_{d_k \in D_{ref}(i)} KPPref_m^{d_k} \quad (13)$$

Using this new score of keyphrases, we can then calculate the retrieval score using equation 9.

4 Results

4.1 Dataset and evaluation

We have opted to test and evaluate our methods on 17164 labelled court decisions from Supreme Court of the Slovak Republic with manually annotated keyphrases. We created a set of all the possible keyphrases and collected their definitions, this set numbering 1500 phrases total. We used these definitions as our query for the document retrieval, labelling the documents that were annotated with the specific keyphrase the definition corresponded with as positive examples and those that weren't as negative examples and evaluated our results using the mean average precision metric, arranging them into a table.

BM-25	cos_full	cos_full + BM-25	cos_averaged	cos_max	keyphrase	keyphrase+ref
78.06	52.77	74.45	69.34	69.55	80.25	92.85

Table 1. The mean average score of our approaches compared to baseline methods.

In Table 1, we compare the mean average precision score of two of our approaches along with 5 baseline algorithms:

1. **BM-25** using the phrases from the query;
2. **cos_ful** cosine similarity between the query and the full document;
3. **cos_ful + BM-25** a metric combining the previous two approaches, given equal weight;
4. **cos_averaged** cosine similarity between the query and its individual chunks after splitting;
5. **cos_max** cosine similarity between the query and its most similar chunk;
6. **keyphrase** Keyphrase based search method without using references described in 3.5;
7. **keyphrase+ref** Keyphrase based search method that uses references described in 3.6.

As we can see from these results, the worst performance belongs to the cosine similarity between the query and the entire document text. We believe this is because judicial decisions are long texts containing many formal parts irrelevant to a given case, causing the embeddings to become muddled and lose their sensitivity. We performed slightly better when we split the document into segments and then calculated the average distance to said segments. Here, the relevant parts separated into specific chunks can increase the average. It seems as though whether we compare the average cosine similarity of the chunk embeddings with that of the query does not yield a significant difference. Still, taking all the other generic texts into account is counterproductive, so if we calculate its similarity to the best matching chunk, we see another performance improvement. Still, we can see that we can only go so far, relying solely on cosine similarity.

Surprisingly, the best baseline algorithm is BM-25, albeit that may be partially due to some of the specific standardized legal terms that were present in both the query and retrieved documents. The second-best results of the baseline algorithms are achieved by combining BM-25 with the entire document cosine similarity, but this is attributed to BM-25's specific suitability. The cosine similarity is more of a hindrance, as BM-25 achieves better results without it. Finally, our proposed methods achieved the best performance, with the one that uses the references outperforming the one that does not by a significant margin. Still, both offer a significant improvement over the baseline algorithms, even if at the cost of additional time and complexity. However this time and complexity is only an issue upon initial calculation and after such calculation is performed for newly inserted documents, queries are real-time with comparable retrieval speeds to most baseline algorithms.

4.2 Ablation study of our algorithm

To determine if and how much said complexity was necessary, we analyzed the performance of our algorithm when we omit certain parts and organized the results into Table 2. Here, we use the same task and the same metric of performance. The columns of the table concern the base keyphrase calculation method from an individual text.

There are six columns, but they are combinations of a few attributes. The first uses either exact matches or similarity matches. Using **exact matches** means that when we evaluate whether a given term is to be taken into account within a given text, we require that the term be present explicitly in the text. In this route, we replace PP_i^t in (1) with the simplified one, where the given term is given a value of 1 if it is explicitly present and given a value of 0 otherwise, ignoring any semantic adjacency to other terms. When we use a so-called **similarity match**, we use PP_i^t in (1) as described, giving a non-zero score to terms semantically similar to our term. These two alternatives are then combined with three possibilities for extracting keyphrases. The first is only using our modified **TF-IDF** for extracting keyphrases, as described in equation (2), leaving out the Attention-seeker score described in equation (5). The columns labeled **Attention-seeker** are the polar opposite, leaving out the calculation of the modified TF-IDF described in equation (2), using only the Attention-seeker algorithm described in (5). Finally, the columns labeled **combined score** use the combination of both of these methods as described in (7).

As for the rows, there are seven in total, but they are also made up of different combinations. The first variant **no references** means that we didn't use any of the extracted references and merely calculated the similarity score using 3.5 or one of its modifications and didn't perform any steps described in 3.6. Rows labeled **legal regulation references** only use the references to legal regulations and ignore referenced court decisions. In comparison, rows labeled **court decision references** take into account only referenced court decisions and not the references to legal regulations. Finally, rows labeled **all references** use both references to court decisions and legal regulations. The other modification concerns whether we calculate the IDF scores for the references themselves or not.

As we can see from this table, most of the extra steps we took offer some performance improvement, and the very best results are achieved when we utilize all the steps described in our algorithm. Looking at the columns, it becomes evident that the combined metric to extract keyphrases is the most effective. Considering not just exact matches but similar phrases, albeit with a reduced relevance, is generally beneficial. We can also see that out of the two algorithms that contribute to the combined score, Attention-seeker produces better results, which is to be expected, considering the simple nature and limitations of TF-IDF. However, it seems omitting it altogether wouldn't yield better results, as the highest score still belongs to the combined metric

If we look at the rows, we can notice that, in general, including references and calculating the IDF scores of the references is rather beneficial. The only

	TF-IDF (exact match)	TF-IDF (similarity match)	Attention-seeker (exact match)	Attention-seeker (similarity match)	combined score (exact match)	combined score (similarity match)
no reference	65.25	68.18	72.91	75.30	80.08	80.25
legal regulation references (no IDF)	67.26	69.97	76.04	79.29	84.39	87.30
legal regulation references + IDF	69.27	74.57	76.37	84.12	88.47	90.58
court decision references (no IDF)	65.73	68.48	73.71	78.60	81.92	82.26
court decision references + IDF	65.95	69.46	75.17	84.09	88.34	88.71
all references (no IDF)	69.54	70.97	75.53	80.65	81.98	86.14
all references + IDF	69.78	75.31	79.81	84.77	89.72	92.85

Table 2. Ablation study of our algorithm

arguably poorly performing metric is relying solely on court decision references, especially when we don't combine it with an IDF score. Depending on the column, sometimes it even offers worse performance than having no reference. This is likely because many cited decisions have more to do with a formal process, and we don't know which part of the decision is cited. Compared to that, the references to legal regulations seem much more useful, although combining legal regulation references and court decision references produces the best results.

5 Conclusion and Future Work

This paper presents an adjustable method that improves baseline document retrieval methods using keyphrase extraction methods. Our preliminary results suggest it could produce better results than many usual approaches. Our best approach, combining parameters from the texts of extracted references, obtained a mean average precision score of 92.85% on our evaluation task. Still, even the version not using the references achieved a performance of 80.25%. We also offer several ways to omit steps from our algorithm in case we don't have the necessary data for references or some of the steps are too computationally complex. In our future work, we aim to create a dataset to test our methods to retrieve the most relevant chunks instead of entire documents, a more difficult and relevant task in Information retrieval. We also wish to test it with more robust language models and evaluate our ranking of these documents against a reranking performed by an LLM.

Acknowledgments

This research for this paper was funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I05-03-V02-00049.

References

1. Hu, W., Zhao, S., Zhao, Q., Sun, H., Hu, X., Guo, R., Li, Y., Cui, Y., Ma, L.: Bert_lf: A similar case retrieval method based on legal facts. Wireless Communications and Mobile Computing **2022**(1), 2511147 (2022)
2. Novotná, T.: Document similarity of czech supreme court decisions. Masaryk University Journal of Law and Technology **14**(1), 105–122 (2020)
3. Pikuliak, M., Grivalský, Š., Konôpka, M., Blšták, M., Tamajka, M., Bachratý, V., Šimko, M., Balážik, P., Trnka, M., Uhlárik, F.: Slovakbert: Slovak masked language model. In: Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 7156–7168. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.530, https://doi.org/10.18653/v1/2022.findings-emnlp.530
4. Pipitone, N., Alami, G.H.: Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. arXiv preprint arXiv:2408.10343 (2024)
5. Szoplák, Z., Gurský, P., Varga, D.: Optimizing Keyphrase Extraction for Court Decisions Using Legal References (12 2024). https://doi.org/10.3233/FAIA241265
6. Tang, L., Clematide, S.: Searching for legal documents at paragraph level: Automating label generation and use of an extended attention mask for boosting neural models of semantic similarity. In: Proceedings of the Natural Legal Language Processing Workshop 2021, NLLP@EMNLP 2021, Punta Cana, Dominican Republic, November 10, 2021. pp. 114–122. Association for Computational Linguistics (2021), https://aclanthology.org/2021.nllp-1.12
7. Undavia, S., Meyers, A., Ortega, J.: A comparative study of classifying legal documents with neural networks. In: Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, Poznań, Poland, September 9-12, 2018. Annals of Computer Science and Information Systems, vol. 15, pp. 515–522 (2018). https://doi.org/10.15439/2018F227, https://doi.org/10.15439/2018F227
8. Varga, D., Gojdič, M., Szoplák, Z., Gurský, P., Horvát, Š., Krajči, S., Antoni, L.: Extraction of legal references from court decisions. In: Proceedings of the 23rd Conference Information Technologies - Applications and Theory (ITAT 2023), Tatranské Matliare, Slovakia, September 22-26, 2023. CEUR Workshop Proceedings, vol. 3498, pp. 89–95. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3498/paper11.pdf
9. Vitório, D., Souza, E., Martins, L., da Silva, N.F., de Carvalho, A.C.P.d.L., Oliveira, A.L., de Andrade, F.E.: Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the brazilian chamber of deputies. Language Resources and Evaluation pp. 1–21 (2024)
10. Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A., Fleisch, B.: Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In: International Conference on Case-Based Reasoning. pp. 445–460. Springer (2024)
11. Z., E.D.L., Tang, C., Shimada, A.: Attention-seeker: Dynamic self-attention scoring for unsupervised keyphrase extraction (2024), https://arxiv.org/abs/2409.10907

Author Index

ISBN 978-4-915905-97-1