CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Extraction of Legal Named Entities from Court Decisions by using Retrieval Augmented Generation and Large Language Models

by

Nayab Sajid

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2024

I dedicate my dissertation work to my parents, supervisor, and all other teachers. A special feeling of gratitude is for my mother, the most responsive woman, I ever know in this world.

CERTIFICATE OF APPROVAL

Extraction of Legal Named Entities from Court Decisions by using Retrieval Augmented Generation and Large Language Models

by

Nayab Sajid

(MCS221008)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Munir Ahmed	BIIT, Rawalpindi
(b)	Internal Examiner	Dr. Aamer Nadeem	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Abdul Qadir	CUST, Islamabad

Dr. Muhammad Abdul Qadir
Thesis Supervisor
November, 2024

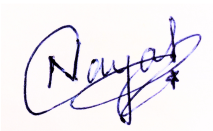
Dr. Abdul Basit Siddiqui
Head
Dept. of Computer Science
November, 2024

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
November, 2024

Author's Declaration

I, **Nayab Sajid** hereby state that my MS thesis titled "**Extraction of Legal Named Entities from Court Decisions by using Retrieval Augmented Generation and Large Language Models**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.



(Nayab Sajid)

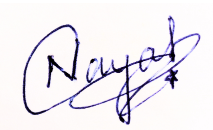
Registration No: MCS221008

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Extraction of Legal Named Entities from Court Decisions by using Retrieval Augmented Generation and Large Language Models**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Nayab Sajid)

Registration No: MCS221008

Acknowledgement

I would like to express my deepest gratitude to Almighty Allah, whose unwavering support and guidance have been the foundation of my journey throughout this thesis. Through every challenge and moment of uncertainty, I have felt His presence and strength guiding me. His wisdom and grace have been a constant source of inspiration and perseverance, making this achievement possible. I would like to extend my heartfelt thanks to everyone who has supported and guided me in achieving this milestone. My deepest gratitude goes to Dr. M. Abdul Qadir for shaping the direction of my research, offering invaluable guidance during moments of uncertainty, and encouraging me.

My family has been a major source of strength and encouragement, providing crucial support whenever I needed it. I am deeply thankful to my beloved parents, spouse and siblings for their unwavering encouragement, patience and prayers which have been my greatest source of strength throughout my MS studies.

The academic development I have achieved is largely due to the dedication and support of the professors at Capital University of Science and Technology (CUST), for which I am immensely grateful. Finally, I would like to express my thanks and offer my regards to all those who have supported me in any capacity.

(Nayab Sajid)

Abstract

Domain-specific Entity Recognition plays a crucial role in the legal domain. Recent advancements have underscored the effectiveness of Large Language Models (LLMs) in natural language processing tasks, showing their ability to accurately detect and classify domain-specific information from specialized texts such as clinical and financial records. This study explores the use of LLMs in identifying domain-specific entities (e.g., courts, petitioners, judges, lawyers, respondents, FIR numbers) within case law documents, particularly focusing on their capacity to handle domain-specific language intricacies and contextual nuances.

This thesis introduces a Hybrid approach to Named Entity Recognition (NER) for Legal Entities, with a particular focus on the InLegalBert dataset. The study identifies limitations in existing models that impact their adaptability and scalability. The existing models that are being used for named entity recognition task are dataset and format dependent, means hundreds of judgments need to be converted into some format which is understandable by these existing models. But if the if the terminologies and nature of expressions of different judgments changed, these model's performance will be significantly decreases. Through a comprehensive analysis, we apply advanced LLMs, like Mistral, for NER tasks, incorporating prompt engineering few-shot learning, and Retrieval Augmented Generation (RAG) techniques which is an emerging methodology in AI.

Our proposed system achieved a notable improvement over the baseline model, reaching an F1 Score of 0.68. The hybrid system, combining prompt engineering and RAG, demonstrated strong precision and straightforward implementation, surpassing existing methods that generally require large datasets and substantial computational resources. These findings suggest that hybrid systems utilizing LLMs, RAG, and prompt engineering can effectively address specific NLP tasks. Named entity recognition using this methodology will not be dataset dependent simple pdf documents can also be used, so it saves the time and cost for data set preparation. we can achieve desired results from the model by making comprehensive prompts from the judgment pdf file.

Contents

List of Figures

List of Tables

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CUST	Capital University of Science and Technology
DL	Deep Learning
GPT	Generative Pre-trained Transformer
HMM	Hidden Markov Model
KG	Knowledge Graph
LLM	Large Language Model
LSTMs	Long Short-Term Memory networks
ML	Machine Learning
MLMs	Masked Language Models
NER	named-entity recognition
NLP	Natural Language Processing
NLU	Natural Language Understanding
QA	Question Answering
RAG	Retrieval Augmented Generation
RNNs	Recurrent Neural Networks
RQs	Research Questions
Regex	Regular Expressions

Chapter 1

Introduction

Entity Identification (EI) is a critical technique within the field of Natural Language Processing (NLP) that serves the purpose of extracting meaningful data from textual content [1]. The primary objective of EI is to locate, recognize, and categorize distinct pieces of information within a text, referred to as" recognized entities." These entities represent the central elements or subjects of the text, which may include a wide range of entities such as the names of individuals, geographic locations, corporate entities, events, products, as well as specific subjects like dates, time periods, themes, monetary amounts, and various other numerical or proportional data.

NER, also commonly referred to by alternative terms such as entity extraction or chunk identification. NER plays an instrumental role in various fields of artificial intelligence (AI), including machine learning (ML), deep learning (DL), and neural network systems. It is an integral part of NLP processes and is widely used in numerous applications, such as search engines, automated sentiment analysis tools, and other data processing systems.

NER task is also vital in other fields for example, in bio-informatics, recognizing names of proteins or genes is crucial. There are ongoing efforts to make such an extended NER's applications span across a variety of industries. It is frequently employed in healthcare for analyzing medical records, in the finance sector for extracting financial data, in human resources for sorting resumes or identifying key competencies, in customer support for automating and improving responses, and in higher education for academic research and data analysis. In densely populated countries like India, China, and Pakistan, the backlog of unresolved court cases continues to grow at an alarming rate [2]. This surge in pending cases has created a pressing demand for systems that can automatically process, interpret, and organize legal documents. One of the key challenges in managing these documents is their unstructured format, extensive length, and the use of complex legal jargon, all of which complicate their processing. The sheer volume of legal documentation, coupled with the intricacies of legal language, requires advanced solutions to efficiently extract, analyze, and interpret the information they contain.

Our aim is to find a way to make adaptable and scalable system which can work on other subdomains of legal domain and numerous NLP tasks like entity recognition, relation extraction, knowledge graph population, question answering system and text summarization can be done using it.

1.1 Legal Named Entity Recognition

Legal Named Entity Recognition (LNER) is a critical subtask within the broader field of NLP that concentrates on identifying distinct entities in legal texts. These entities typically consist of important components such as geographic locations, organizational names, and individual persons, all of which hold semantic relevance in legal documents [3]. Legal texts are often dense and filled with jargon, which makes the process of entity recognition crucial for breaking down the information in a structured manner. LNER's role is to pinpoint these entities and their relationships within documents, making it easier to process, analyze, and retrieve information from vast legal records. Unlike general NER tasks, where the focus might be on commonly recognized entities like names of people or companies.

The overarching goal of LNER is to develop sophisticated models that not only recognize these legal entities but also establish the connections and relationships between them, which forms the foundation for relationship extraction. This process involves understanding how various entities interact within a legal document such as how a petitioner relates to a court, or how a statute connects to a legal case. Once these entities and their relationships are identified, they can be used to populate knowledge graphs, which serve as structured repositories of information. These knowledge graphs are then applied in various legal AI systems, enabling them to provide more informed and accurate responses to questions posed in the legal domain. In this way, LNER supports advanced legal research tools, legal analytics, and automated decision-making systems.

Entity Recognition, when customized for specific domains like law, plays a pivotal role in enhancing numerous legal applications. For instance, in legal contexts, LNER supports the automatic summarization of complex legal texts, enabling users to quickly grasp the essential points without reading lengthy documents. It also underpins question-answering systems that can respond to legal queries by extracting relevant information from documents, as well as sentiment analysis systems that assess the tone or implications of legal discourse. Furthermore, LNER aids in the retrieval of critical information from legal documents, which is essential for lawyers, researchers, and legal professionals who need quick access to pertinent data. By identifying key legal entities and relationships, LNER helps to organize legal information in a way that makes it easier to navigate and interpret. This is particularly useful in law-related contexts, where the volume of unstructured data can be overwhelming.

Recent advancements in NLP have highlighted the growing success of LLMs in performing complex tasks like entity recognition, especially in specialized domains such as law, healthcare, and finance [4]. These models have shown great promise in identifying and categorizing specialized information, such as recognizing legal entities within documents that contain dense and complex language. For instance, LLMs can be fine-tuned to identify specific legal entities such as courts, petitioners, judges, lawyers, respondents, and even First Information Report (FIR) numbers in case law documents. By leveraging their ability to understand context and semantics, LLMs are able to navigate the complexities of legal language, providing accurate and relevant information extraction.

Following is an example taken from the website of Supreme court that shows the structure of the judgment document. We are interested to find legal entities from judgment documents.

IN THE SUPREME COURT OF PAKISTAN
(Original Jurisdiction)

PRESENT:
Mr. Justice Umar Ata Bandial, CJ
Mr. Justice Ijaz ul Ahsan
Mr. Justice Sayyed Mazahar Ali Akbar Naqvi
Mr. Justice Jamal Khan Mandokhail
Mr. Justice Muhammad Ali Mazhar

SUO MOTO CASE NO.3 OF 2022
(Re: Independent and Transparent Investigation into the Murder of Renowned Journalist, Mr. Arshad Sharif in Kenya).

In attendance:

Ch. Aamir Rehman, Addl. AGP
Dr. Akbar Nasir, IG ICT
Mr. Awais Ahmad, DIG
Mian Shahbaz, IO
Mr. Irfan, Director Law M/o Information
Mr. Israr Ahmad Khan, Director Law FIA
Mr. Waqas Rasool, Deputy Director Law
Mr. M. Syrus Sajjad Qazi, Addl. Secretary M/o Foreign Affairs
Mr. Murad Wazir, DG M/o Foreign Affairs
Mr. Asad Khan Burki, LA M/o Foreign Affairs
Syed Faraz Raza, ALA M/o Foreign Affairs

Date of Hearing: 05.01.2023

FIGURE 1.1: An example of named entity recognition from wikipedia

The source of the example is available on the web, given in the Figure 1.1, it is an example of preamble section of named entities that contain entities organization, location, date, person, weapon. Entity recognition tailored to specific domains is essential in natural language processing, especially in fields like law. It involves identifying and categorizing key judicial entities such as petitioners, respondents, judges, and attorneys, which are central to various applications. These applications comprise extraction of relations, machine translation, sentiment analysis, faceted search, knowledge base development, and information retrieval, as highlighted by Thomas and Sangeetha [5].

ORDER

The learned Addl. Attorney General has referred to the report filed by the Special Joint Investigation Team (**"SJIT"**) on 04.01.2023 which recounts the progress made so far in its investigative work since constitution of the SJIT on 07.12.2022. He has informed us that the Federal Government has provided requisite funds for investigative work to be done by the SJIT in two foreign countries, namely, UAE and Kenya. Requests for Mutual Legal Assistance (**"MLA"**) to the government in these countries were issued by the Federal Government on 04.01.2023. In this behalf he acknowledges that the Foreign Office is cooperating fully to secure the requisite help and collaboration of

FIGURE 1.2: An example of named entity recognition from wikipedia

The source of the example is available on the web, given in the Figure 1.2, it is an example of judgment section of named entities mentioned above. Recognizing specific entities within a domain and understanding how they relate to each other not only improves the indexing and retrieval of legal documents but also acts as a key step in selecting features for tasks like clustering, classification, and extracting information for summarization [6]. An effectively designed entity recognition (ER) system places a strong foundation for several applications in the legal division, including the following:

Some applications of NER are:

Legal Question-Answering Systems: Judicial facts play a critical role in determining answers to fact-based queries. For instance, in response to a query like," Who is the appellant in a particular judgment?" the answer is likely to be a judicial entity. If the response is" Mr. Ali" the dataset is annotated with judicial entities, the question-answering model recognize"Mr. Ali" as an entity, potentially identifying it as the correct solution.

Knowledge Graph Creation: By identifying named entities (NEs) in judicial texts and understanding the relationships between them, we can represent textual data graphically, such as through entity relationship graphs. These graphs are valuable for answering complex relational queries and facilitate text summarization by highlighting and annotating the most relevant information associated with a NE.

Case-Based Reasoning: Case-based reasoning relies on the knowledge derived from information extracted from court language. This extracted information can be utilized by various proficient systems.

Relation Extraction (RE): Relation extraction relies heavily on entity recognition in judicial documents, as it helps identify key entities like judges, plaintiffs, defendants, legal bodies, and locations referred to in the text. Once these entities are recognized, relation extraction can determine the connections between them, such as 'defendant charged with a crime,' 'plaintiff initiated a lawsuit against defendant,' or 'court ruled in favor of the plaintiff.

Additionally, the relation triplets generated through this process can be used as features in various applications of machine learning, such as classification of text, summarization of documents, and paraphrase identification.

1.2 Types of Legal Named Entity Recognition

1.2.1 Rule-based System

The rule-based method is considered a precursor to modern NLP techniques. This approach involves using predefined linguistic rules to process and analyze text data.

One of the primary benefits of rule-based systems is their interpretability, as well as their ability to capture patterns specific to a given domain. Experts in the field create these rules manually to reflect syntactic structures and linguistic phenomena relevant to the target corpus. This makes it easier to tailor the system for specific domains. Often, rules are implemented using regular expressions (regex), which offer a concise and flexible way to match text patterns.

Rule-based system can be used for development of NER, which is actively used in tasks such as machine translation, information extraction, automatic summarization, and question answering systems. NER is a crucial component in these applications, as it helps identify and extract relevant entities from text data. It enables accurate processing and understanding of the content, which is essential for effective translation, information extraction, summarization, and answering user queries [7].

Rule-based approaches come with several key advantages in NLP. Their interpretability stands out, as the rules are human readable and easy to understand, making system behavior transparent. Additionally, these systems often out perform statistical or machine learning models in terms of processing speed for certain tasks. Their deterministic nature also offers a high level of control, allowing for precise system adjustments by modifying the rules as needed.

However, rule-based approaches are not without their limitations. Developing and maintaining a comprehensive set of rules requires significant linguistic expertise. While these systems can achieve high precision, they often suffer from lower recall, meaning they may fail to identify patterns not covered by the predefined rules. Moreover they are not scalable to the other subdomains or if dataset is changed because terminologies may change in other datasets.

1.2.2 Machine Learning (ML)

Machine learning models, especially in the realm of NER, have historically relied on conventional statistical techniques such as linear regression, logistic regression, and decision trees. These approaches require the manual design of features, which involves identifying the key attributes from the input data that are believed to be most relevant for the task at hand. In the case of NER, these features might include lexical characteristics like word shape, capitalization, prefixes, suffixes, part-of-speech tags, or surrounding word context.

Various Machine Learning techniques, such as support vector machines (SVMs) and Hidden Markov models (HMMs), have been employed in the literature. The NLP framework includes modules for syntactic processing, including tokenization, part-of-speech tagging, and sentence detection. Additionally, the NLP systems incorporate modules for named entity recognition, relation extraction, and concept identification [8].

For tasks such as NER, which involves identifying and classifying entities like people, organizations, and locations within text, these traditional models can yield reasonable performance with relatively modest datasets. This is especially true when the entity categories are simple and well defined. In fact, by carefully crafting features that highlight the linguistic patterns associated with named entities, these models can be effective for smaller-scale projects where computational efficiency and simplicity are prioritized [9]

However, despite their potential to perform well on certain NER tasks, these models are inherently constrained by the quality and relevance of the manually engineered features. The success of such models largely hinges on the expertise of the developer to accurately capture the nuances of the data. This can be especially challenging in NER, where the contextual and semantic complexity of language plays a critical role in determining the boundaries and classifications of entities.

Furthermore, these limitations become even more apparent when working with more complex or domain-specific NER tasks, such as in legal or biomedical texts, where the entities are often highly specialized and less predictable. In such cases, feature engineering can become a labor-intensive process that may not fully capture the intricacies of entity types and their context within the document. The model's reliance on predefined features may also hinder its adaptability to new, unseen datasets, making it less effective in dynamic environments where language and entity usage evolve over time.

Therefore, while traditional machine learning models offer a foundation for NER tasks, their performance is closely linked to the quality of the features designed, and they often fall short when tasked with handling the complexity and variability inherent in real-world language[10].

1.2.3 Deep Learning (DL)

Deep learning approaches have revolutionized the field of NER by offering highly effective solutions for capturing the complexities inherent in natural language. These methods excel at automatic feature extraction, allowing models to uncover intricate patterns and relationships in the data without the need for manual feature engineering. This is especially important in NER tasks, where the goal is to identify and classify entities such as names of people, organizations, locations, or specific terms within a wide variety of texts, ranging from news articles to legal documents.

Deep learning has revolutionized many application domains, ranging from image processing to NLP. Deep learning based solutions to NER task has turned out to be very common recently and are achieving state-of-the-art results. The two most common deep learning models for NER are Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [11]. Over the past decade, DL has achieved great success due to its powerful capability in automatically learning complex features of data. DL models excel at extracting meaningful patterns and representations from large and diverse datasets, which has enabled significant advancements in various fields, such as computer vision, natural language processing, and speech recognition [12].

One of the key advancements in NER has been the fine-tuning of pre-trained language models such as BERT. These models, which have been trained on massive amounts of text data, provide a solid foundation for recognizing entities in a variety of contexts. Fine-tuning involves adapting these pretrained models to the specific task of NER by using task-specific labeled data. This approach allows the models to transfer the rich linguistic knowledge they have already acquired to the NER task, significantly improving accuracy and efficiency.

A key advantage of deep learning-based NER models is their ability to capture complex semantic and logical relationships within the text. Unlike traditional machine learning methods that depend on manually engineered features, deep learning models can automatically learn these relationships, enabling them to handle a wide range of entity types, even in complex or ambiguous scenarios. For instance, in legal documents, where the terminology and entity categories can be highly specialized, models like BERT can be fine-tuned to recognize domain specific entities without requiring extensive manual feature design.

In summary, deep learning approaches have provided state-of-the-art (SOTA) solutions for NER by enabling models to automatically learn the complex dependencies and relationships in text. Techniques like BERT, GPT, and LSTMs have proven to be highly successful in extracting the rich, contextual information necessary for accurately identifying and classifying named entities. Their ability to generalize across different datasets, coupled with their superior handling of complex language structures, has made deep learning the go-to method for modern NER tasks [4]. But these models can not evolve themselves according to the terminologies and changes occurring in the field, if there information is not provided in the dataset there performance will not be up to mark.

1.2.3.1 Large Language Models

There are other type of models that can be used to perform NLP tasks like named entity recognition. These models are trained on billions of parameters, they sore the changes occurring in everyday language, they store information about the evolutions occurring in terminologies and expressions of everyday languages. LLMs perform well when prompt engineering is used along. Prompt engineering with large language models adds another layer of sophistication to hybrid systems.

This technique involves crafting specific prompts or queries to guide the model's responses or behavior. In the context of NER, prompt engineering can be used to elicit more accurate and contextually relevant responses from large language models, improving their ability to extract and classify entities effectively. For example, carefully designed prompts can help the model focus on identifying specific types of entities or understanding the nuances of entity relationships.

In large language models (LLMs), prompts constructed based on the in-context learning method play a crucial role in instructing the models. This method primarily provides LLMs with a small number of input-output pairs, and then instructs the models to perform specific tasks based on these few-shot demonstrations. Furthermore, as the demonstration of these few-shot examples can be considered a form of fine-tuning for the model, the selection of the demonstrations is of paramount importance [13].

By integrating prompt engineering with large language models into NER architectures, hybrid systems can achieve significant improvements in performance. The combination leverages the strengths of deep learning in handling complex and dynamic language tasks, while also utilizing the precise control provided by prompt engineering and the contextual knowledge from other methodologies. This integration allows for more accurate and efficient NER, capable of handling a wide range of texts and entity types with enhanced precision and contextual understanding.

In summary, LLMs in NER represent a cutting-edge approach that combines the capabilities of deep learning with traditional methods and advanced techniques like prompt engineering. This integration enhances the model's ability to manage complex, ambiguous, and context-sensitive entity recognition tasks, leading to improved overall performance, scalability and adaptability in diverse applications.

1.2.4 Hybrid system

Hybrid systems in Named Entity Recognition represent an advanced approach that integrates multiple methodologies to leverage their respective strengths. These systems combine deep learning with other techniques such as traditional machine learning, rule-based methods, and ontology-based systems, aiming to enhance overall NER performance.

Deep learning models, such as BERT and GPT, have revolutionized NER by providing sophisticated tools for understanding context and semantic relationships within text. These models excel in handling complex, ambiguous, and dynamic language tasks, thanks to their ability to automatically learn from large datasets and capture nuanced patterns in text.

On the other hand, traditional machine learning methods, rule-based systems, and ontology-based approaches offer distinct advantages. Machine learning models rely on engineered features and can be tuned for specific tasks with domain-specific data. Rule-based systems use predefined linguistic rules to identify entities, providing precision in well-defined contexts.

Combining these methodologies can address the limitations of each individual approach. For instance, integrating deep learning models with rule-based systems can provide the robustness of deep learning in handling complex patterns, while also benefiting from the precision and specificity of rule-based approaches. Similarly, incorporating ontology-based methods can enhance the contextual understanding of deep learning models by providing structured knowledge about entity relationships. Together they can work well but still they are not able to adapt themselves according to the changes occurring in some domain's terminologies, nature of expressions and new terminologies and conditions that are added to the domain (i.e Legal, Medical, Finance).

1.3 Problem Formulation with Significance

NER system for legal domain is of utmost importance. Previously the systems used for LNER task do not store information for the evolutions occurring in the terminologies and expressions of legal domain. A probable Solution is to used Large Language Models that are trained on natural languages and then use special constructs in the prevailing legal for further training. LLMs can be made scalable and adaptable to the legal subdomains such as criminal, medical, corporate, court rulings etc. using prompt engineering methods. But previously some of the models used large language models for NER task but their performance is still not good (there may be a need to fix the context in an effective way) [6]. So, our aim is to find a way that can enhance the performance of LLM in the context of legal documents

1.4 Research Questions

Following are the Research Questions that need to be addressed:

- RQ1: How can we optimize the LLM to make it adaptable for domain specific tasks?
- RQ2: How comprehensive prompts can be designed to extract named entities from legal documents by using LLM?

1.5 Research Objectives

Following research objectives (ROs) need to be answered:

- RO1: Optimizing the LLM to make it adaptable for domain specific tasks
- RO2: Designing comprehensive prompts to extract named entities from legal documents by using LLM?

Chapter 2

Literature Review

2.1 Introduction

Research in Natural Language Processing within legal named entity recognition advancing rapidly, leveraging the vast amounts of unstructured data available in judgment documents. By integrating NLP with other AI-driven approaches, the goal is to find an architecture that is adaptable to other subdomains of legal domain and is scalable to other NLP tasks like question answering systems, text summarization etc. Various researchers have proposed techniques for Named Entity Recognition. The insertion criteria of our research for choosing literature review papers were based on their significance to Named Entity Recognition methods like rule-based methods, model fine tuning of pretrained models and large language models. Their paper achieved an F1 score of approximately 91%. These models perform best on the dataset upon which they are trained but when the dataset changed and new terminologies are used there performance decreases.

Modi et al. [14] develop a legal NER model using a dataset of 14,444 judgment sentences of Indian court and 2,126 preambles, that is annotated for 14 legal entities. They create a transformer-based baseline model for legal entity recognition and employ rule-based post-processing to address document-level context and co-reference resolution. Their dataset includes 11,970 judgments from 29 Indian courts, representing a mix of high court and supreme court cases.

The study compares two NER architecture types trained on this dataset. The first uses a transition-based dependency parser on top of a transformer model, while the second involves fine-tuning the transformer with an added linear layer. For the transition-based parser, they experimented with the Roberta-base model and InLegalBERT using the Spacy library, achieving an F1 score of 91.1%. The fine-tuning approach was tested with Roberta-base, InLegalBERT, and legalBERT using the TNER library.There paper achieved f1 score of 90.99% by using technique Ensemble model with dynamic weights 90.99%.

2.2 Comparative Analysis of Existing Techniques

The analysis of existing literature reveals that the majority of the Natural Language Processing systems can be categorized into Deep Learning, Rule-based, Hybrid approaches and Large Language Models. Majority of work is being done on deep learning techniques especially on transformer models.

2.2.1 Deep Learning

2.2.1.1 Transformer Model (Bert based)

Dhannachendra et al [15] to capture the unique semantics of legal documents, developed a large language model (LLM) specifically for Indian legal documents, based on RoBERTa, with a focus on masked language modeling. They used this LLM to train two different NER models: a generic NER model and a legal NER model. The generic NER model was further fine-tuned on the CoNLL-2003 dataset, which includes entities like persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC).

This model was used in post-processing to detect entities that the legal NER model missed. For the Spacy legal NER model, they fine-tuned the legal RoBERTa-base model on the CoNLL-2003 dataset. To enhance the model's understanding of legal semantics, they used the updated weights from the generic NER model and further trained it with the LNER dataset using Spacy transformers. After obtaining predictions, they applied rule-based post-processing to address entity recognition issues.

Additionally, they proposed two baseline models. The first, Legal LM-NER, involves fine-tuning the legal LLM using Hugging Face transformers. The second model combines the contextual word embeddings from their LLM with the prediction outputs of the Kalamkar et al. (2022a) model, passing them through a classification layer.

Quang Minh et al. [16] evaluate various benchmark pre-trained language models like BERT, RoBERTa, and InLegalBERT for NER tasks. They then combine the outputs of these PLMs in an ensemble module to extract the final results.

They handled the benchmark dataset using two different approaches. The first approach uses a model that is based on the framework of spaCy, which employs a transition-based parser on top of a PLM. The PLMs used include RoBERTa, BERT, LegalBERT, InLegalBERT, and XLM-RoBERTa, with all experiments conducted within the spaCy framework.

The second approach involves using the IOB format, combined with dependency parsing and a CRF layer, to classify each word in a sequence. Various PLMs such as XLM-RoBERTa, RoBERTa, DeBERTa, BERT, LegalBERT, InLegalBERT, and InCaseLegalBERT were selected as backbones for this approach.To improve performance, they applied an ensemble strategy, joining models through majority voting with weights. For each token, the final label was predicted based on the majority label of the models. The f1 score of their paper was 0.90%.

The model's performance on the private test set was suboptimal in the second approach. To enhance its effectiveness, it is necessary to fine-tune the weights assigned to each Pretrained Language Model (PLM) within the ensemble module. This optimization is expected to improve overall performance.

Intisar et al. [17] developed a model based on BERT, focusing on improving the process of named entity recognition in the legal domain. In the preprocessing stage, they initially converted the dataset from spaCy's JSON format to the BIO (Beginning, Inside, Outside) format to ensure compatibility with the model. The data was then tokenized using the BertTokenizerFast class from the transformer's library, specifically utilizing the LEGAL-BERT model introduced by Chalkidis et al. in 2020. To maintain consistency between the original labels and the tokenized sentences, they adopted the alignment method proposed by Winastwan (2022), ensuring that the word IDs matched accurately.

For the training process, the pretrained LEGAL-BERT model served as the foundation for classifying text at the token level. They employed the BertForToken Classification class for this task. The model was trained using a standard PyTorch training loop, where hyperparameters were fine-tuned based on the model's performance on a development dataset. They achieved f1 score of 87%.

Yuri et al. [18] conducted an in-depth exploration of using transformer models such as RoBERTa, InLegalBERT, and XLNet for Named Entity Recognition tasks. The goal was to identify and classify legal entities like petitioners, respondents, courts, or statutes. During preprocessing, they applied the BIO (Beginning, Inside, Outside) labeling format and employed a sliding window technique to prevent input truncation. This method was only triggered when the length of the text exceeded the maximum input capacity of the transformer model.

The baseline model followed a standard token classification approach. For more advanced custom models, the researchers experimented with two variations, building on the best-performing transformers from the baseline. Specifically, they used RoBERTa combined with BiLSTM and CRF layers, based on the work of Dai et al. (2019), and XLNet with BiLSTM and CRF layers, drawing from both Rongen Yan (2021) and Dai et al. (2019). These models were implemented using Py Torch, with custom classes created to integrate BiLSTM and CRF components.

While the BiLSTM layer was part of the PyTorch library, the CRF layer was implemented using the pytorch crf5 library. The CRF layer utilized the Viterbi Algorithm, enabling it to calculate loss and predict the most probable labels for the text.

The training process was carried out using the transformers library's Trainer class, specifically designed for RoBERTa-based models. To further enhance performance, the team developed a label remapping function based on Regex patterns, which allowed them to align the predictions more accurately with the original text. This adjustment resulted in an improvement in RoBERTa's baseline performance.

Among the models tested, XLNet emerged as the top performer, outperforming both InLegalBERT and RoBERTa in both baseline and custom configurations. However, a key limitation identified in their model was that it could not process entire documents but was instead restricted to handling individual sentences, which posed a constraint in certain legal tasks requiring broader document context.

Jingjing et al. [19] outlined a methodology where they initially employed continued pretraining using the LegalBert20w model within a Bert-CRF framework, which acted as their baseline model for the task. They observed a slight improvement when they introduced prototypical contrastive loss to the model, enhancing its performance marginally. For their final submission, the team implemented an ensemble of models, each trained on different datasets and multiple auxiliary tasks. This ensemble approach further improved the overall system's effectiveness, demonstrating the benefit of combining various models and tasks for enhanced performance. They achieved an f1 score of 86%.

Irene et al. [3] explore the use of Natural Language Processing (NLP) techniques to support lawyers and domain experts in retrieving content and making decisions within the legal sector. Their study focuses on LNER and employs a recently developed entity-aware attention mechanism, implemented through the LUKE model (Yamada et al., 2020). This model generates contextualized representations for both entities and words concurrently.

The research involves a comprehensive analysis of transformer-based models for L-NER, with a particular emphasis on the impact of domain-specific pretraining and the effectiveness of fine-tuning these models. Their objective is to evaluate how pretraining influences the quality of model representations, the efficacy of transfer learning in the legal domain, and the capability of pre-trained models to grasp legal language and specialized terminology.

The study assesses a range of transformer-based models, including those utilizing entity-aware methods designed to focus on domain-specific entities. It also provides an in-depth comparison of various explainable AI techniques and introduces a novel approach to NER within the legal context.This approach aims to enhance the accuracy and relevance of entity recognition by integrating advanced techniques tailored to the legal field. They achieved an f1 score of 83%.

Ginn Khamov et al. [20] discuss their approach for SemEval-2023 Task 6, Subtask B, which involves NER in legal documents, specifically targeting various types of legal entities. The legal documents are categorized into two sections: preamble and judgment texts, with certain entities being tagged only in one of these sections. To tackle this challenge, the team proposed a to- ken classification model that incorporates information about the document type, resulting in enhanced performance compared to a system that did not use this augmentation.

The research began with training and evaluating a baseline system, alongside several variations, through four distinct experiments. These models were trained with parameters from Kalamkar et al. (2022). In addition to the baseline, the team tested LegalBERT in place of RoBERTa, increased the dropout rate from 0.1 to 0.3, and expanded the hidden state size from 64 to 128. They trained five custom models using fine-tuned transformers, initially starting with two models based on LegalBERT and RoBERTa. RoBERTa was ultimately selected as the superior model for further augmentation.

Three distinct augmentation strategies were evaluated: one strategy involved adding the sentence class as an input token, another used a twin model where the sentence class determined between two sub-models, and the third applied the sentence class in the final linear layer. Each model was trained on the complete training dataset and validated on a subset, using 40 epochs of training with a batch size of 64 and 3 gradient accumulation steps. They achieved an f1 score of 72%.

The training process employed the Adam W optimizer (Loshchilov and Hutter, 2017), with a learning rate of 2E-5, weight decay of 0.01, beta1 set to 0.9, beta2 set to 0.999, and epsilon of 1E-8. Training was conducted on an Nvidia RTX A6000 GPU, with each training session taking around 3 hours.

Yimin et al. [21] have introduced a new linguistic resource, a corpus of 19th-century Estonian Parish Court records, annotated specifically for named entities. Although the texts were manually digitized, they display significant variability in spelling conventions, capitalization, and dialects. While their named entity taxonomy extends beyond the typical categories such as Person, Location, and Organization, they demonstrated that it is still possible to achieve high inter-annotator agreement. This annotated dataset was utilized to fine-tune and evaluate several BERT-like transfer learning models.

Despite all models being pretrained on contemporary Estonian texts, their performance on historical Parish Court records was unexpectedly strong. One possible reason for the high NER quality is that manual digitization resulted in clean, error-free text. Additionally, the relatively uniform structure of Parish Court records may have further contributed to this success.

Nida et al. [22] conducted an investigation into the effectiveness of BERT models for extracting legal entities from court documents, focusing on several variations of the BERT architecture. Their study aimed to assess how well these models could handle legal texts by using two distinct datasets of court judgments from Pakistan. Specifically, they employed datasets from the Lahore High Court and the Supreme Court of Pakistan, both of which contain rich legal information.

In addition to these datasets, Nida et al. manually labeled a specialized dataset consisting of 214 Civil Appeal judgments from the Supreme Court of Pakistan.

This dataset was annotated with fourteen distinct entity types relevant to legal proceedings. This comprehensive labeling effort aimed to provide a robust training foundation for evaluating the models.

The team fine-tuned three BERT variations—BERTBASE-cased, BERTBASE-uncased, and LegalBERT on their annotated dataset to determine which model achieved the best performance in terms of F1 score, a metric that balances precision and recall. The results indicated that the BERTBASE-cased model out- performed the other models, achieving the highest F1 score.

This suggests that BERTBASE-cased was particularly well-suited for capturing the nuances of legal language and extracting entities effectively. In contrast, the BERTBASE-uncased and LegalBERT models, while still effective, yielded slightly lower F1 scores. The performance of the BERTBASE-cased model was notably superior compared to previously reported results, highlighting its potential for improved accuracy in legal entity extraction tasks. This indicates that, for the specific context of legal documents and the datasets used, the cased variant of BERT provides a more refined and accurate approach for identifying and extracting legal entities.

Nida et al. study demonstrates the potential of fine-tuning BERT models for specialized legal tasks, providing valuable insights into the effectiveness of different BERT variations in handling legal texts and suggesting that the BERTBASE-cased model is a particularly strong candidate for future research and applications in legal entity extraction.

2.2.1.2 Transformer model (Large Language Model)

Gupta et al. [4] trained an NER model using OntoNotes v5 (Ralph Weischedel et al., 2011) and applied it to the documents of legal domain using a zero-shot learning approach. Meanwhile, Kalamkar et al. developed an NER corpus derived from various Indian court judgments and trained an NER system using a Transformer-based architecture (Vaswani et al., 2017), incorporating coreference resolution and rules to harmonize named entities.

Atin et al. [6] conducted a performance assessment of several advanced LLMs, including LLaMA 3 (Large Language Model Meta AI 3), Mistral, and Gemma, with a focus on entity recognition (ER) tasks within Indian legal texts. Their study employed few-shot prompt engineering techniques to harness the capabilities of these models for judicial entity recognition in legal documents.

In their approach, they designed carefully crafted prompts to guide the LLMs in generating outputs formatted according to predefined JSON specifications. This method was aimed at optimizing the models' performance in identifying and classifying legal entities accurately. The evaluation involved analyzing the effectiveness of several state-of-the-art LLMs, such as LLaMA 3, Gemma, Mistral, and Phi 3, in the context of legal text processing.

To measure the models' performance, Atin et al [6] used precision, recall, and F1 score as key evaluation metrics. These metrics provided a thorough assessment of each model's ability to accurately identify and label entities within the legal documents. Precision measures the accuracy of the identified entities, recall assesses the model's ability to find all relevant entities, and the F1 score offers a balanced view of the model's overall performance in entity recognition tasks. This comprehensive evaluation aimed to highlight the strengths and weaknesses of the various LLMs in the context of judicial entity recognition, offering insights into their effectiveness and potential applications in the legal domain. They achieved an f1 score of 63% by using large language model Mistral.

Bhaskarjit et al. [23] highlight the difficulties LLMs face in extracting and interpreting complex information from unstructured legal documents, despite employing advanced techniques like Retrieval Augmented Generation (RAG), particularly VectorRAG methods that use vector databases for information retrieval. These challenges are primarily attributed to the specialized terminology and the complex structure of legal documents. To address these challenges, we present a new approach called that combines RAG, Mistral and prompt engineering. This approach improves question-answer (QA) systems, named entity recognitions systems, text summarization systems for extracting information from legal documents, generating precise and contextually relevant answers.

Jionglong et al. [19] present a novel training algorithm that harnesses the power of RAG to refine and enhance the performance of LLMs in specialized tasks such as headline generation and numerical value extraction from unstructured data. The proposed methodology centers on identifying the most relevant training examples that serve as contextual references for fine-tuning the LLMs, ultimately leading to improved accuracy and context-sensitive outputs. The core of their approach involves the integration of extended markup language (XML) tags specifically designed to improve the model's ability to handle and interpret numerical information.

During the preprocessing phase, numbers em- bedded within headlines are enclosed in XML tags before being input into the model. This tagging technique serves a dual purpose: it clearly delineates nu- numerical data from the rest of the text, and it aids the model in recognizing and processing these numbers with greater precision. To further support the model's ability to manage numerical data, the training process incorporates a variety of mathematical operations. This feature is integral to the chain-of-thought (COT) reasoning process, which helps the model perform complex computations and derive accurate results. By referencing and executing different mathematical operations, the model can engage in a structured reasoning process that enhances its capability to solve mathematical problems effectively.

Following the model's generation of headlines or numerical data, a comprehensive post-processing phase is implemented to ensure the accuracy of the outputs. This validation step involves cross-checking and confirming the correctness of the numerical values produced by the model, thereby ensuring that the results meet high standards of accuracy and reliability. The application of this advanced retrieval-augmented generation technique, coupled with XML tagging and COT reasoning, led to substantial improvements in the model's performance.

The systematic approach to handling numerical data and the rigorous validation of results resulted in notable enhancements in the quality of generated headlines. These advancements were reflected in superior performance evaluations during human assessments, demonstrating the effectiveness of the proposed algorithm in outperforming alternative methods.By integrating RAG with XML tagging and COT reasoning, Jionglong et al. have developed a robust framework that not only enhances the model's ability to generate and process numerical data but also ensures that the results are accurate and contextually appropriate. This innovative approach represents a significant advancement in the field of LLM training and application, offering promising improvements for tasks involving complex data extraction and generation.

Yucheng et al. [24] address the ongoing challenges faced by LLMs in the field of NLP, specifically issues like hallucinations (generation of incorrect or misleading information) and the requirement for domain-specific knowledge. Despite the significant advancements made by LLMs, these challenges persist, necessitating improvements in their functionality and reliability.

To tackle these issues, recent strategies have focused on integrating external information retrieval mechanisms with LLMs. This combination has been shown to greatly enhance the performance of LLMs across a variety of NLP tasks by supplementing the models' internal capabilities with relevant external data.

The paper delves into the essential components of Retrieval-Augmented Language Models (RALMs), which include Retrieval-Augmented Generation (RAG) and Retrieval-Augmented Understanding (RAU), as well as other augmentation techniques. RAG involves augmenting the generation process with retrieved information, allowing the model to generate responses that are informed by additional relevant data. RAU, on the other hand, enhances the model's understanding by incorporating external retrieval into the comprehension process.

Yucheng et al. also examine how these components interact to form different model architectures and applications. RALMs are versatile and can be applied to a broad spectrum of NLP tasks, including but not limited to translation, dialogue systems, and knowledge-intensive applications. By leveraging external information, RALMs are able to produce more accurate, contextually relevant outputs and handle complex queries that require specialized knowledge. Overall, the integration of retrieval mechanisms with LLMs represents a significant advancement in NLP, addressing some of the limitations inherent in traditional LLMs and enabling more effective and nuanced performance across a range of tasks.

Soto-Jimenez et al. [25] highlight the growing importance of question-answering (QA) systems in recent years as essential tools for combating misinformation and providing accurate information in response to user queries. In their research, they introduce the design of a QA system that is built upon a RAG framework. This innovative approach combines the strengths of semantic retrieval models, which are used to extract relevant information from vast textual datasets, with the power of pretrained language models to generate precise and contextually appropriate answers.

The system operates by retrieving documents from a continuously updated database containing domain-specific content and feeding them into the generative module, which then produces reliable and informative responses. A key advantage of this design is its inherent flexibility; the system's dataset can be dynamically constructed and updated by drawing information from Wikipedia pages, allowing it to be easily adapted to different fields of knowledge.

To validate the efficacy of this approach, the authors developed a QA prototype with a focus on the pollution domain. This prototype underwent three iterations of development, during which it successfully addressed critical challenges, such as improving operational efficiency and enhancing the quality of the responses generated. The preliminary testing of the prototype yielded positive results, showing that the RAG-based system performs effectively. Notably, only 10 percent of the responses exhibited minor inconsistencies with the context after evaluating a small set of queries. The findings of study suggest that this RAG-based framework has significant potential for developing robust and reliable QA systems. Its adaptability across various fields, coupled with the continuous updating of domain-specific content, positions it as a promising tool for delivering precise and context-aware answers. Additionally, the ability to dynamically build and update the dataset using in- formation from sources like Wikipedia further enhances the system's versatility, making it suitable for a wide range of applications where accurate information retrieval and generation are essential. These models tend to achieve F1 scores ranging from 70% to 85% depending on the quality of embeddings and specific tuning processes applied to retrieval and answer generation stages.

Yan Hu et al. [26] explore the significant potential of advanced language models, such as GPT-3.5 and GPT-4, in processing complex clinical data and extracting meaningful insights with minimal reliance on extensive training datasets. By developing and fine-tuning prompt-based methods, these models demonstrate improved performance, making them highly practical tools for NER tasks, while also reducing the need for large-scale annotated data.

The study specifically evaluates the effectiveness of GPT-3.5 and GPT-4 in handling clinical NER tasks, proposing task-specific prompts to enhance the models' accuracy. The research focuses on two primary clinical NER tasks: (1) extracting medical concepts, such as medical problems, treatments, and tests, from clinical notes in the MTSamples corpus, a dataset aligned with the 2010 i2b2 concept extraction shared task, and (2) identifying adverse events related to nervous system disorders from reports in the Vaccine Adverse Event Reporting System (VAERS). To optimize the performance of these models, the researchers introduced a structured, task-specific prompt framework.

This prompt framework was designed around several key components: (1) baseline prompts, which provided an overview of the task and specified the format of responses; (2) prompts crafted using detailed annotation guidelines; (3) prompts developed from insights gathered through error analysis; and (4) few-shot learning prompts that leveraged a small set of annotated examples to guide the model's output. Each of these prompts was systematically evaluated for its effectiveness in enhancing the model's NER capabilities. Additionally, GPT-3.5 and GPT-4 were bench marked against BioClinicalBERT, a well-known model specialized in biomedical text processing.

The results highlight the adaptability of GPT-3.5 and GPT-4 to clinical NER tasks, demonstrating their potential in scenarios where annotated data is scarce. By using tailored prompts and incorporating a structured framework, the study shows how these models can perform on par with, or even surpass, specialized models like BioClinicalBERT in extracting valuable medical insights. This research underscores the potential of prompt-based approaches in making state-of-the-art language models more accessible and efficient for clinical applications, offering a promising direction for future work in the domain of healthcare and clinical data processing.

Shiye et al. [27] highlight the growing importance of extracting hidden legal insights from large collections of legal documents, a task that has gained both academic and practical relevance due to the increasing transparency of judicial information. While LLMs have achieved remarkable success in many NLP tasks, particularly in text generation, obtaining high accuracy in specialized tasks like legal entity relation extraction remains a significant challenge. This difficulty stems mainly from the limited availability of well-annotated legal data and the laborious, time-consuming process required to label such datasets. As a result, research in few-shot learning has become more prominent in this field.

In this study, the authors introduce a few-shot learning approach for entity relation extraction within the legal domain, utilizing the capabilities of large pre-trained models. These models, drawing on their extensive prior knowledge, can quickly adapt to new tasks with minimal annotated data. The proposed method was tested on two publicly available datasets focused on legal entity relation extraction. The experimental results demonstrate that this approach not only significantly reduces the cost and effort involved in creating training data but also achieves impressive performance in few-shot legal entity relation extraction tasks.

By leveraging few-shot learning, the study provides a solution to the scarcity of annotated legal data, showing how LLMs can be effectively applied in specialized legal tasks. This method allows for rapid adaptation to new legal extraction tasks with minimal training, making it a promising approach for legal tech applications where annotated data is often scarce and costly to produce. The results affirm the potential of few-shot learning in improving accuracy and efficiency in legal entity relation extraction, offering a viable path forward for advancing legal NLP research and practice.

Gustavo et al. [28] emphasize the significant role of information extraction in the legal domain, particularly in situations where vast amounts of unstructured data, such as legal opinions, are available but structured, machine-readable data is limited. When properly processed, these unstructured legal documents can yield valuable insights into previous lawsuits, thereby helping legal professionals make more informed decisions and enabling the development of data-driven applications. This paper specifically addresses information extraction within the Brazilian legal system, focusing on structuring features from legal opinions related to consumer complaints.

To achieve this, the authors explore two distinct approaches. The first approach employs traditional supervised learning techniques, treating the extraction of categorical features as a text classification task, while numerical feature extraction is accomplished through NER. This conventional method relies on well-annotated training data to accurately extract relevant information from legal texts.

The second approach takes advantage of the increasing prominence of LLMs, utilizing ChatGPT and prompt engineering to extract both categorical and numerical features. By leveraging the flexibility and generalization capabilities of LLMs, this method offers an alternative that potentially reduces the reliance on large annotated datasets while still delivering high-quality information extraction. Through these two methods, the paper investigates the effectiveness of combining traditional supervised learning techniques with cutting-edge LLMs like ChatGPT.

The research highlights how these approaches can be applied to the Brazilian legal context, providing structured insights from legal opinions on consumer complaints. This study underscores the potential of both methods to improve the accessibility and utility of unstructured legal data, offering valuable tools for legal professionals and researchers aiming to enhance data-driven legal practices. The findings show that while both approaches deliver similar results based on traditional evaluation metrics, ChatGPT significantly simplifies the process and reduces the time needed for extraction.

Fabian et al. [29] discuss the potential of large language models (LLMs) to generate high-quality, human-like text and emphasize their application in named entity recognition (NER), a crucial task in natural language processing (NLP) that focuses on identifying relevant entities within text documents. In their paper, they introduce "llmNER", a Python library specifically designed to facilitate zero- shot and few-shot NER using LLMs. The library features a user-friendly interface that allows users to easily create prompts, query the LLM, and process the responses generated. One of its main strengths is that it simplifies prompt engineering, offering a streamlined interface for experimenting with different parameters to optimize the NER performance. The development of llmNER aims to make the process of prompt creation and response parsing more accessible, enabling researchers and practitioners to conduct experiments in in-context learning with greater ease.

The software was validated across two different NER tasks, demonstrating its versatility and effectiveness in performing both zero-shot and few-shot learning. By offering a simplified approach to prompt engineering and LLM querying, llmNER advances research in NER and promotes further exploration in in-context learning by minimizing the complexities involved in using large language models for entity extraction tasks.

Bonifacio et al. [30] stated that Deep language models such as ELMo, BERT, and GPT have revolutionized NLP by achieving remarkable results across various tasks, including NER. These models typically follow a two-phase training process. First, they undergo pretraining on large, general-domain, unlabeled corpora, allowing them to learn a wide range of language structures and patterns. After this general pretraining, the models are fine-tuned in a supervised manner on specific downstream tasks such as NER, sentiment analysis, or machine translation. However, a less explored and optional step in this process involves further fine-tuning the pretrained language models on an intra domain corpus, which consists of unlabeled text from the same domain as the target task, before the final supervised training. This step, which could provide substantial gains for domain-specific tasks, remains underutilized and understudied in existing literature.

In this work, they address the gap in understanding the impact of this intermediate fine-tuning phase, particularly in the context of Named Entity Recognition for Portuguese legal documents. Legal texts are known for their complexity and use of specialized terminology, which often differs significantly from general-language texts. As a result, models trained solely on general-domain corpora may struggle to accurately identify entities within legal documents. To address this, we evaluate the effect of domain-specific fine tuning on NER performance using two prominent deep learning architectures: ELMo (Embeddings from Language Models) and BERT (Bidirectional Encoder Representations from Transformers). Additionally, they experiment with four distinct unlabeled corpora, each relevant to the legal domain, and assess the models across three Portuguese legal NER tasks.

Their experimental results reveal that finetuning language models on a large, domain-specific corpus of legal texts leads to significant improvements in NER performance for legal documents. This highlights the importance of domain relevance when training language models for specialized tasks. By exposing the models to legal texts during the finetuning phase, they are better equipped to handle the unique terminology, sentence structures, and contextual subtleties found in legal documents. Consequently, the models can more effectively recognize legal entities such as case numbers, court names, statutes, and other domain-specific terms, resulting in higher accuracy and improved NER outcomes.

To investigate whether these improvements stemmed solely from the increased exposure to domain-specific data or whether they were due to the additional training on legal texts, we also evaluated the performance of the finetuned models on two general-domain NER tasks [31] . Interestingly, the results showed that while finetuning on legal corpora led to significant gains in legal NER tasks, it actually resulted in a performance decline on general-domain tasks. This finding suggests that domain- specific finetuning enhances the model's ability to handle specialized content at the expense of its generalizability. In other words, the model becomes highly op- timized for the legal domain but loses some of its effectiveness when applied to more general NER tasks.

One of the standout findings from their study is that the BERT model, finetuned on a legal domain corpus, significantly outperforms the previous state-of-the-art on the LeNER-Br corpus. LeNER-Br is a Portuguese-language NER corpus specifically designed for the legal domain, and achieving such improvements demonstrates the tangible benefits of domain-specific finetuning. This breakthrough confirms that BERT's contextualized representations, when tailored to the legal domain, are highly effective at capturing the intricacies of legal language and improving entity recognition in legal documents.

Moreover, the success of this finetuning approach provides important insights into the future of NER for specialized domains. It shows that while pretraining on general-domain data offers a strong foundation, additional domain-specific finetuning can unlock further potential, especially in areas like law, medicine, or finance, where language usage differs markedly from everyday texts. Finetuning allows the model to learn domain-specific patterns, making it better suited to tasks that require a deep understanding of specialized vocabulary and entity relationships. In addition to enhancing NER performance in legal texts, their research highlights the broader implications for other domain-specific tasks. For instance, similar approaches could be applied to fields such as medical NER, where entities like diseases, treatments, and pharmaceuticals must be identified accurately. In the medical domain, finetuning models on clinical or biomedical corpora could lead to better performance in extracting critical information from patient records, research papers, and other specialized documents.

This study underscores the importance of finetuning deep language models on intradomain corpora for improving the performance of NER systems in specialized fields like law. While general-domain pretraining provides a solid foundation, domain-specific finetuning enables models to better capture the linguistic nuances of legal texts, resulting in significant improvements in legal NER tasks [32] . However, this approach comes with trade-offs, as the increased specialization may reduce the model's performance on general-domain tasks. Nonetheless, the ability to optimize NER systems for specific domains offers promising directions for future research, particularly in developing more effective NLP solutions for fields that require precise entity recognition, such as law, medicine, and finance.

Serena et al. [33] reviews developments in AI applied to legal contexts over the past decade, marking a transition from symbolic approaches to those primarily using machine learning. This shift emphasizes the application of natural language processing in legal data management and the broadening of AI's capabilities in categorizing and analyzing legal documents. The paper reviews eight significant studies, including work on web-based legal document management and applications of machine learning in various legal areas, underscoring the increased focus on using language models to address complex legal tasks.

B.Samrah et al. [23] introduces a combined RAG system that merges Vector-RAG and GraphRAG methodologies to enhance data retrieval for information extraction. This model integrates a vector-based retrieval system, which captures similarity-based information, with a knowledge graph that organizes structured relationships, offering more comprehensive context for generating responses.

By uniting these approaches, HybridRAG provides responses that are both contextually detailed and relevant to structured data. Evaluations show that HybridRAG outperforms separate RAG methods in terms of accuracy, completeness, and relevance of responses, highlighting its utility in tasks that require both unstructured and structured data sources.

S.Tallam et al. [26] presents a solution for headline generation and numerical reasoning, key elements of SemEval-2024 Task 7. This approach implements a Retrieval-Augmented Generation (RAG) model with three main components: a knowledge base, a dense retrieval mechanism, and a headline generator. Using pre-trained BERT embeddings, the knowledge base enables the model to retrieve news articles with contextual similarities, assisting in the generation of accurate headlines.

To optimize headline accuracy, the model uses the LLAMA2-7b architecture, which applies the retrieved content iteratively to refine token predictions. This RAG-based model achieves strong results on ROUGE metrics, with notable improvements in recognizing numerical data and creating structured headlines, showcasing the model's potential for tasks involving both text generation and numerical precision. Notably, they achieve a ROUGE-1 score of 48.08 with their end-to-end training.

Y.Yang et al. [34] proposes a semi-supervised tri-training method for named entity recognition (NER) tailored to legal cases. It aims to improve entity identification accuracy within knowledge bases of criminal case properties, addressing data scarcity challenges through multiple classifiers. By enhancing entity recognition in legal texts, this approach is designed to aid law enforcement with more accurate data integration in legal applications.

J.huo et al. [35] addresses the complexities involved in interpreting legal documents, which often feature lengthy passages, specialized vocabulary, and a lack of sufficient annotated datasets. To tackle these challenges, the authors developed a hierarchical approach that employs domain-specific pretraining alongside data augmentation and auxiliary-task learning techniques. Additionally, they integrated an ensemble method to bolster system performance. This strategy led to a first-place finish in the rhetorical role classification sub-task (RR) and achieved commendable results in other related sub-tasks.

2.2.2 Machine Learning Systems

Luis et al. [36] present an NER system trained using Frustratingly Easy Domain Adaptation (FEDA) across manifold legal corpora. Their primary goal was to create an NER capable of detecting 14 types of legal entities in Indian judgments. Besides the FEDA architecture, they explored using overlapping context and averaging tensors to process lengthy legal texts, which proved advantageous.

The FEDA architecture is described as a collection of dense layers built at top a pretrained language model. It includes a general FEDA layer and several specialized FEDA layers, each connected to the general layer. The FEDA layers are defined and linked in a specific manner to train the model. Specifically, a FEDA layer F consists of a GELU activation layer followed by a Linear layer. A General FEDA layer G is a stack of two F layers joined by a Layer Normalization layer. Each Specialized FEDA layer Cx consists of two F layers connected by a Concatenation layer and a Layer Normalization layer, with the G layer linked to each Cx layer through their respective concatenation layers. Additionally, a Conditional Random Field (CRF) is placed atop all the specialized Cx layers to enhance the NER system. In summary, the G layer acts as a classifier that processes inputs from all datasets, while each specialized classifier focuses on entries specific to its corresponding dataset. The CRF layer is shared across all datasets.

All names of entities were encoded using the BILOU labeling scheme. The authors used DeBERTa V3 (He et al., 2021) Large as the pre-trained language model and trained multiple models during the evaluation period, maintaining consistent hyperparameters except for the sequence size in DeBERTa.

Sometimes their system had problems predicting a party correctly, either Petitioner or Respondent, when it was a person representing an organization. This was hard to train, not only because of the particularity of the guidelines but also because of the span of the context that had to be executed to recognize the entity accurately. For the issues they have faced for getting good results they should create a exclusive CRF for different dataset, that should also speed up the training of a model.

Can Cetindag et al. [37] states that NLP technologies are rapidly advancing in the legal domain, with a growing focus on automating and improving tasks such as document analysis, legal research, and contract review. One of the most important

NLP tasks in this field is NER, which involves identifying and classifying named entities such as people, organizations, laws, and locations within legal texts. In the context of legal NLP, NER plays a crucial role due to the extensive use of specific entities and terminologies in legal documents. However, despite its importance, domain- specific NER models tailored to the legal field remain under explored, particularly for languages other than English.

This study addresses this gap by presenting a NER model specifically developed for Turkish legal texts, supported by a custom-built corpus. The research introduces several NER architectures that leverage both traditional machine learning and deep learning approaches, including CRFs and BiLSTMs.

To enhance the performance of these NER models, the study [38] investigates various word representation techniques. Specifically, it explores the use of GloVe embeddings, Morph2Vec, and neural network-based character-level feature extraction methods. These techniques are critical for capturing the nuances of Turkish, an agglutinative language where words often consist of multiple morphemes that modify meaning.

The most notable result of this study is the achievement of promising F1 score using a hybrid word representation approach, combining GloVe and Morph2Vec embeddings with character-level features extracted through BiLSTM. This result underscores the effectiveness of integrating multiple embedding techniques and deep learning architectures to enhance NER performance, particularly in a domain-specific and morphologically rich language like Turkish.

Moreover, the study acknowledges the unique challenges posed by Turkish as an agglutinative language, where words can take on many forms depending on their grammatical context. This linguistic characteristic makes it more difficult for standard NER models to accurately identify entities without incorporating morphological features. By addressing this, the study not only contributes to the development of legal NER for Turkish but also provides insights that could be applied to other agglutinative languages, such as Finnish, Hungarian, and Korean.

In addition to its significance for Turkish legal NLP, this research is pioneering in that it is the first study to develop a domain-specific NER model for the Turkish legal field. It also stands as one of the first efforts to tackle NER in an agglutinative language within the legal domain. The implications of this work extend beyond Turkish, offering a framework that could be adapted for legal NER in other languages with complex morphological structures.

Hossein et al. [39] stated that NER plays a pivotal role in numerous high-impact applications such as knowledge base creation and semantic search, as it automates the identification of entities like people, organizations, and dates from text. Despite significant advancements in developing machine learning models for generic NER tasks, these models often fall short when applied to specialized domains such as law or finance, where domain-specific entities are critical. Generic models typically excel at recognizing common named entities but struggle to identify nuanced legal or financial entities like case numbers, statutes, or regulatory terms, which require more contextual understanding. Consequently, tackling domain-specific NER necessitates the creation of ground truth data and experimentation with different models to achieve the desired performance.

This paper focuses on the challenges and solutions to NER in the legal domain, specifically motivated by a real-world use case in the financial sector. In specialized fields like law, producing high-quality labeled data is crucial, yet this process is particularly difficult due to the expertise required for accurate annotation. Legal texts often contain a wealth of technical jargon, formal language, and entity types that are vastly different from those found in general texts. As such, models trained on generic NER tasks do not adapt well to the legal context without domain-specific data.

To address this issue, they propose a comprehensive approach for solving NER in the legal domain, starting with the construction of a large dataset of legal documents through web crawling. A semi-automated labeling process is then implemented to generate high-quality ground-truth labels for a set of eleven predefined legal entities. These entities include key legal concepts such as court names, case numbers, legal statutes, and other terms specific to legal documentation. The semi-automated labeling process ensures consistency and accuracy while reducing the manual burden on human annotators.

The proposed approach significantly improves the quality of the labeled data, surpassing popular off-the-shelf NER tools that are not tailored for the legal domain. By incorporating human expertise with automated processes, the method delivers labels that accurately capture the intricate relationships and meanings of legal entities. Moreover, this semi-automated pipeline can be applied to generate ground-truth labels for an unlimited number of legal documents, allowing for scalability across different legal cases and jurisdictions.

Following the creation of high-quality labeled data, they experiment with various NER models and training methodologies, fine-tuning them for the legal domain. The models range from traditional machine learning approaches to deep learning models like CRF, BiLSTM-CRF, and transformer- based architectures such as BERT. Each model is trained using the generated legal dataset and evaluated on the predefined legal entities.

Additionally, the dataset, labels produced by human annotators, and the semi-supervised labeling method are made available to the research community, along with the code used for training and evaluation. This transparency and resource-sharing will enable further research in domain-specific NER tasks and allow other researchers to replicate or extend the work for different legal systems or specialized domains.

In conclusion, this paper highlights the challenges of domain-specific NER in the legal field, emphasizing the necessity of generating high-quality ground-truth labels and fine-tuning models to adapt to the unique requirements of the domain. The results demonstrate that a semi-automated approach, combined with state-of-the-art models, can significantly improve NER performance in specialized fields. This approach not only provides valuable insights for NER in legal documents but also opens the door for further advancements in NER across various specialized domains such as medicine, finance, and academia.

Peter et al. [40] introduces a new class of neural network architectures specifically designed for graph-structured data. These Graph Attention Networks (GATs) employ masked self-attention mechanisms to address the limitations associated with traditional graph convolutional methods. By allowing nodes to selectively attend to the features of neighboring nodes, GATs can dynamically assign different levels of importance to these connections without relying on complex matrix computations or prior knowledge of the graph structure.

This innovative approach tackles several challenges inherent to spectral-based graph neural networks and is applicable to both inductive and transductive learning tasks. The performance of GATs has been validated against several established benchmarks, achieving or surpassing state-of-the-art results on datasets like Cora, Citeseer, and Pubmed, as well as a protein-protein interaction dataset.

Wang et al. [41] presents a specialized approach for processing legal texts, addressing tasks like legal NER and rhetorical role classification. The researchers implemented a legal-contextualized model using the LUKE framework, which combines a bidirectional transformer architecture with specific token and entity embeddings fine-tuned for legal language contexts.Their methodology was tested on two primary tasks: identifying specific legal entities within texts, such as "petitioner," "respondent," and "court," and classifying rhetorical roles, with the goal of enhancing interpretability in legal documents.

The Legal-LUKE model demonstrated notable improvements over traditional BERT-based methods, especially in managing the unique complexities and terminology of legal language. These findings underscore the model's potential to improve both accuracy and efficiency in automated legal text processing tasks, ultimately aiding legal professionals and researchers in managing complex legal documentation.

Bosch et al. [42] showcases the contributions of students from the University of Orléans who participated in SemEval Task 6, which focuses on enhancing the efficiency of legal professionals through automated tools. The team addressed two specific sub-tasks: Legal Named Entity Recognition (LNER) and Rhetorical Role

(RR) prediction, with the goal of creating systems that are both lightweight and easy to interpret.

For the LNER task, the researchers implemented a CRF model, which achieved a macro F1 score of 0.74 on the development set and 0.64 on the evaluation set, while also incorporating post-processing techniques for certain named entities.In the RR task, they developed two different classification systems—one based on the Bag-of-Words approach and another utilizing a sentence-transformer model.

The Bag-of-Words method, which integrated results from the LNER system, performed better, achieving a macro F1 score of 0.49 on the development set and 0.57 on the evaluation set.

2.2.3 Hybrid Systems

Pin Tang et al. [38] present a sophisticated deep learning methodology known as BiLSTM-CRF, which integrates BiLSTM networks with CRF to enhance NER in legal texts. The BiLSTM component of the model captures contextual information from both preceding and succeeding words in a sequence, while the CRF layer optimizes the sequence of labels, ensuring that the predicted entity tags follow logical patterns. The study utilized a dataset of 1,000 legal judgment documents, which were obtained from China Judgments Online, a comprehensive repository of legal documents. The dataset was meticulously divided into three parts: 600 documents for training the model, 200 for validation during the training process, and 200 for final testing to evaluate model performance. This division ensured that the model was properly trained, fine-tuned, and tested on distinct sets of data to avoid overfitting and to provide a reliable measure of its effectiveness.

During preprocessing, the documents were standardized by removing extraneous spaces to ensure uniformity. The team applied BIO labeling using the YDEEA corpus annotation tool, which involves tagging words in the documents with labels that indicate their role within the named entity categories—beginning, inside, or outside of an entity. This standardization step was crucial for training the BiLSTM-CRF model to accurately identify and classify entities.

The authors recommended using the Adaptive Moment Estimation (Adam) optimizer, a popular optimization algorithm that adapts the learning rate based on the moments of the gradient, to enhance the model's performance. Adam's adaptive learning rate helps in achieving better convergence and stability during the training process.To assess the model's effectiveness, it was tested on documents pertaining to specific legal contexts, including commutation, parole, and temporary release from prison. These documents were also sourced from China Judgments Online, ensuring consistency in the data used for evaluation.

That achieved an f1 score of 85.5%. Despite the overall effectiveness of the BiLSTM-CRF model, the performance was notably lower when analyzing the Docket Number field. The challenge with docket numbers arises from their complex structure, which often includes a mixture of numbers, characters, and parentheses. This complexity makes it difficult for the model to achieve consistent and accurate extraction, resulting in lower precision, recall, and F1 scores for this specific field.Overall, the study demonstrates the BiLSTM-CRF model's potential in enhancing NER tasks within legal documents, while also highlighting the challenges posed by certain types of data, such as docket numbers, that require further refinement and specialized handling.

Trias et al. [43] presented an innovative ensemble language model designed to extract names from legal texts in English by combining a transformer-based neural network with a finite state machine. This approach leverages the strengths of both technologies to enhance the accuracy and efficiency of name extraction from legal documents.

The model was trained and evaluated using data sourced from the Harvard Caselaw Access Project, a comprehensive repository of legal documents from the United States. The process involved two key stages: first, the extraction of names using the ensemble model, and second, the application of heuristic textual analysis to address and rectify errors.

In the initial stage, the ensemble model, which integrates the transformer-based neural network architecture with a finite state machine, was employed to extract names from the legal texts. The transformer component of the model is adept at understanding context and semantics, while the finite state machine helps in managing the sequences and patterns associated with names. This combination enhances the model's ability to accurately identify names within the complex structure of legal documents.

Following the extraction, the system applies heuristic textual analysis to identify and address potential errors in the extracted names. This process involves automatic correction of many of the errors and flags any remaining issues for manual review. The heuristic methods are designed to recognize common mistakes and discrepancies, improving the overall accuracy of the name extraction process.

The paper outlines a two-step methodology for the extraction of lawyer names. The first step involves using the ensemble model to extract names from the texts, while the second step focuses on identifying and correcting transcription errors to ensure each lawyer's name is accurately and uniquely identified. They achieved an f1 score of 93%.

Harshil et al. [44] explained that widespread application of BERT, one of the leading language models, has significantly advanced many NLP tasks. One of the key tasks where BERT has shown substantial improvements is NER, which involves the automatic identification of entities like locations, persons, organizations, and more from a given text. NER is crucial as it serves as a foundational step for several downstream NLP applications, including information extraction, argumentation mining, and document classification.

Despite the extensive research focused on applying BERT and other popular language models to NER in general, the use of BERT for Legal NLP specifically in the legal tech domain remains relatively under explored. Legal NLP involves applying specialized NLP techniques, such as sentence similarity and NER, to legal data, which tends to differ significantly from general-domain text due to its complexity, jargon, and highly structured language. While there are models tailored for NER tasks using BERT, few focus on legal documents, and even fewer target non-English legal texts, particularly those in German.

In this paper, they address this gap by fine-tuning a German BERT model on a Legal Entity Recognition (LER) dataset specifically curated for German legal documents. This task is crucial for automating legal processes, such as document review, contract analysis, and case law research, where accurately identifying legal entities like court names, case numbers, statutes, and legal provisions is essential.

To ensure the robustness and generalizability of their model, they employ a stratified 10-fold cross-validation. This approach helps mitigate overfitting and ensures that the model's performance is evaluated comprehensively across different subsets of the data. Cross-validation is particularly important in legal NLP tasks, where the variability in legal language across different cases and document types can be significant.

The results of their fine-tuning experiments show that German BERT, when trained on the LER dataset, significantly outperforms traditional models, including the BiLSTM-CRF+ architecture used by the creators of the LER dataset. This performance boost highlights the advantage of transformer-based models like BERT in capturing the intricate contextual dependencies in legal texts, which are often missed by older, recurrent-based models like BiLSTM.

By leveraging the pretrained German BERT model, we capitalize on its strong understanding of the German language while adapting it to the legal domain, resulting in more accurate identification of legal entities. This advancement is particularly important for improving Legal Tech solutions in German-speaking jurisdictions, where the legal industry is seeking more automation tools for document analysis.

Finally, recognizing the importance of open access and reproducibility in NLP research, they make their fine-tuned German BERT model available on HuggingFace. This will enable other researchers and developers in the Legal Tech community to further explore its applications in legal NER tasks or integrate it into broader legal automation systems. By doing so, they hope to contribute to the growing body of Legal NLP research and encourage more domain-specific NLP solutions tailored to legal systems worldwide. They achieved an f1 score of 92%.

Leitner et al. [45] states that this paper presents a comprehensive approach to NER within the context of German-language legal documents. The focus is on developing a robust NER system specifically for legal texts, which often contain complex language and specialized terminology. To achieve this, a dataset was created using German court decisions, meticulously annotated with 19 semantic classes, including person, judge, lawyer, country, city, street, landscape, organization, company, institution, court, brand, law, ordinance, European legal norm, regulation, contract, court decision, and legal literature. These diverse classes capture the full range of entities present in legal documents, reflecting the unique requirements of the legal domain.

The dataset comprises approximately 67,000 sentences and includes over 54,000 annotated entities. To enhance its utility and adaptability, the original 19 fine-grained entity classes were generalized into seven broader, coarse-grained categories: person, location, organization, legal norm, case-by-case regulation, court decision, and legal literature. This dual-level annotation structure provides flexibility, enabling researchers to choose between fine-grained or coarse-grained annotations depending on the task or use case. The availability of both detailed and simplified entity classifications makes this dataset versatile for various NER tasks, facilitating the training of models that can adapt to different levels of granularity required by specific legal document processing tasks. They achieved an f1 score of 95%.

In the paper [46] For the NER task, the paper explores two families of state-of-the-art models: CRFs and BiLSTMs. CRFs, a traditional machine learning approach known for their effectiveness in sequence labeling tasks like NER, were evaluated alongside BiLSTMs, which are a more advanced deep learning method capable of capturing long-range dependencies in text. Given the complexities inherent in legal documents, which often contain lengthy and intricate sentences, BiLSTM models are particularly well-suited for handling these challenges.

These findings underscore the advantage of using deep learning-based models like BiLSTM over traditional machine learning models like CRFs for legal NER tasks, especially when dealing with the nuanced and specialized language of legal texts. BiLSTMs are better equipped to capture contextual dependencies and entity relationships that are critical in legal documents, where the meaning of a word or phrase can often depend on its broader context.

The research conducted in this paper contributes to the European project LYNX, which aims to develop a semantic platform for document processing and analysis within the legal domain. The successful application of NER in legal documents is essential for automating tasks such as legal information retrieval, contract analysis, and court decision review, ultimately improving the efficiency and accuracy of legal research and documentation.

2.2.4 Rule-based Systems

Raabia et. al. [47] developed CustNER, a rule-based named-entity recognition system that has improved recall compared to other NER models. CustNER uses a set of 7 rules to identify named entities in text. The rules are designed to improve recall by capturing entities that may be missed by machine learning-based NER models. Evaluation results show that CustNER has higher recall compared to other rule-based and machine learning-based NER systems, while maintaining a high level of precision.

The key advantages of the CustNER model are its ability to identify more named entities, particularly rare or uncommon ones, that were missed by other NER models. This makes CustNER a valuable tool for applications that require high recall in named entity recognition, such as information extraction, question answering, and knowledge base population. Additionally, the rule-based approach of CustNER makes it more interpretable and easier to customize for specific domains, compared to black-box machine learning models.

CustNER was able to identify more named entities, particularly rare or uncommon ones, that were missed by other NER models. This is a significant advantage, as accurately recognizing and extracting named entities is crucial for many natural language processing tasks, such as information extraction, question answering, and knowledge base population.

Table 2.1: Literature Review

Ref	F1 Score	Technique	Dataset
[15]	91.20%	Perform best when used RoBERTa-base	InLegalNer
[16]	90.98%	Performed best using RoBERTa+ Transition based parser dynamic weights	InLegalNer
[36]	90.7%	DeBERTa V3+ FEDA+ CRF	InLegalNer
[17]	87.94%	LegalBERT	InLegalNer
[18]	87.43%	XLNet+BiLSTM+CRF	InLegalNer
[19]	86.22%	LegalBERT+CRFensemble	InLegalNer
[3]	83.21%	LUKE-base/BERT-base	InLegalNer
[20]	72.65%	RoBERTa+Sentenceclasstoken	InLegalNer
[41]	66.70%	legal-LUKE	InLegalNer
[42]	64.89%	SpaCy(preprocessing)+CRF	InLegalNer
[48]	55.32%	customtrainedspaCy+BERT-CRF	InLegalNer
[49]	51.73%	ALBERT / RoBERTa + BiLSTM / ID-CNN + CRF	InLegalNer
[29]	54%	Mixtral 8x7B	CONLL-2003
[6]	63%	Used LLM Mistral	InLegalNer
[50]	74%	GPT-3	DODFCorpus I

2.3 Identified Research Gaps

Literature reviews indicate that pretrained and rule-based models are increasingly employed due to their superior performance and flexibility. However, they demand substantial computational resources and time. Moreover, these models are not able to provide information about the revolution occurring in terminologies and nature of expressions in legal documents if that is not included in the dataset used to train that model. So, they may not be used for the subdomain of legal domain and also may not be used to perform other NLP tasks like judgment document summary, question answering systems.

The solution is to use Large Language Models that are trained on billions of parameters and they store the evolutions occurring in the terminologies, nature of expression, ways of writing court decisions and the other laws and regulations that are made day by day.

So large language models are the good choice to resolve the issue of adaptability across various subdomains and scalability across NLP tasks. LLMs also handle tasks with minimal labeled data, performing well with zero-shot or few- shot prompt engineering methods. Some of the papers have used large language models for NER task but their performance is still not good because there is a need to fix the context that a LLM used to generate result for some domain specific task like Legal named entity recognition.

So, our objective is to find a technique that can increase the performance of LLM in the context of legal documents. One of the techniques that can be used to enhance the performance of Large Language Model is to use RAG (Retrieval Augmented Generation) which provide context to the LLMs and facilitate the model to correctly identifying entities and then design comprehensive prompts. These prompts will can be the examples from the dataset which helps the model to extract relevant data or entities from the dataset. This dataset can be any document from which we are interested to extract entities. We do not need to annotate a huge number of judgments in this technique.

Chapter 3

Proposed System - A Hybrid Approach

In this chapter, a practical exploration of the proposed system model and the detailed design of a hybrid system is undertaken. Chapter 3 serves as a direct continuation of the comprehensive study initiated in Chapter 1, where the motivation and overarching objectives of the research were introduced.

1. Entities are from Legal domain using dataset InLegalNer (2022)
2. Previous systems are not Adaptable across subdomains of Legal Domain
3. Previous Systems are not scalable to the other NLP tasks for legal domain
4. This methodology's goal is to evaluate the performance of Large Language Models for the task of legal named entity recognition when RAG and Prompt Engineering is used

In this research, we utilize few-shot prompt engineering to harness the capabilities of large language models for judicial entity recognition (ER) in legal documents. This method entails designing a single, well-structured prompt that guides the LLM to produce responses in a designated JSON format. The JSON output contains both the extracted text and the corresponding entity labels from the input document. This approach is especially beneficial as it eliminates the need for extensive task-specific training. By leveraging the pretrained LLM's sophisticated natural language understanding, we can efficiently detect and label entities within legal texts, simplifying the process and minimizing the effort typically required for model training and fine-tuning.

3.1 Dataset

The InLegalNer [14] is high quality annotated dataset which contained 14 legal entities. Corpus is made up of 14444 Indian judgment sentences of court and 2126 preamble sentences. Figure 3.1 is an example taken from taken from InLegalNer dataset. This Corpus is a representative sample of Indian court judgments that contains 11970 judgments across different 29 Indian courts. The dataset has two parts

- Preamble
- Judgements

The dataset is in Json Format. Below figure 3.2 represents the structure of the dataset used for experimentation.

Considering above mentioned annotated text, below is the demonstration of the Entities used in the dataset.

Court entity is present in both preamble section and in judgment section and this entity is basically the court names present in the current judgment.

Petitioner entity is present in both preamble section and in judgment section and this entity is basically the names of petitioners or appellants present in the current judgment.

Respondent entity is present in both preamble section and in judgment section and this entity is basically the court names of the respondents or defendants or the oppositions present in the current case.

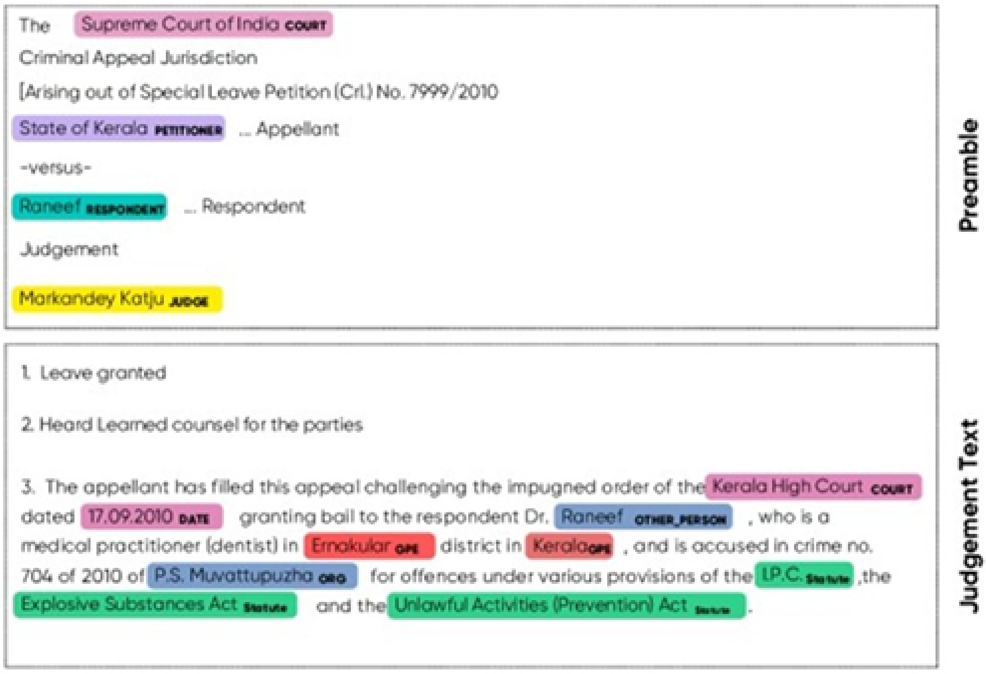


FIGURE 3.1: An example of Legal Named Entities in Court decisions from dataset InLegalNer

Judge entity is present in both preamble section and in judgment section and this entity is basically the name of judge of present and can also be the name from previous judgments.

Lawyer entity present only in the preamble section and it can be the names of the lawyers of both parties.

Date is the entity which can be any date mentioned in the case.

ORG is the entity which can be the name of any organization mentioned in the case and it can be a bank, or any government agency etc. This entity is extracted from the judgment section.

GPE is a geo political entity. It can be any country, city, village, town name that is mentioned in the case and this entity is extracted from the judgment section.

Statute is the entity which is basically name of any law or act that is mentioned in the case and this is present only in judgment section.

Provision is the entity which depicts, articles, orders, Sections, sub-sections, rules under a statute and this also extracted from the judgment section of the corpus.

```
                    ]
                },
                "id": "UMPQI6BX",
                "from_name": "label",
                "to_name": "text",
                "type": "labels"
            },
            {
                "value": {
                    "start": 231,
                    "end": 243,
                    "text": "Shyam Sunder",
                    "labels": [
                        "RESPONDENT"
                    ]
                },
                "id": "IHZKN3AB",
                "from_name": "label",
                "to_name": "text",
                "type": "labels"
            }
        ]
    }
],
"data": {
    "text": "In the case of State of M.P. v. Shyamsunder Trivedi (supra), which was a case of custodial
},
"meta": {
    "source": "criminal_gujarat_high_court judgement https://indiankanoon.org/doc/54915645"
}
```


FIGURE 3.2: JSON structure of dataset

Precedent is the entity that can be the previous references of the cases and also it can be citations or party names. This is extracted from the judgment section.

Case-Number is the entity that basically is all the case numbers that are mentioned in the judgment and it do not contain party names and citations and this also is extracted by the judgment section.

Witness is the entity that can be the names of all the witnesses that are present in the current judgment.

Other_person This entity is basically the names of all other persons that are not the witness or petitioner or respondent or judge.

By understanding the nuances of these entities and their roles in legal documents, we can unlock the full potential of legal text analysis and contribute to the advancement of legal research and decision-making, leading to more efficient and accurate legal processes.

3.2 Proposed System

In this section, we proposed an approach that used LLM Mistral, few shot prompt Engineering method and RAG. Our system processes the entire dataset, ensuring that the results are generated with a comprehensive consideration of both the data and the prompts designed for that data.

The dataset utilized in our methodology is the preprocessed InLegalNer dataset, which has been previously employed by numerous data scientists for various experiments. This dataset, which was also used in the SemEval 2023 Task B, provides a robust foundation for our study. The InLegalNer dataset is known for its relevance in legal domain NER tasks, making it an appropriate choice for evaluating the effectiveness of our proposed method.

For the LNER task, we employed the Large Language Model Mistral, known for its advanced capabilities in handling complex language tasks. Additionally, we utilized RAG to enhance the experiment's depth and effectiveness. Prompts were crafted using the training dataset and subsequently applied to the test dataset.

This approach ensures that the prompts are well-tuned to the specifics of the data, thereby improving the overall accuracy and performance of the system in recognizing and classifying named entities. The prompt engineering approach focuses on a different strategy. This approach involves a detailed understanding of statement structures, crafting effective prompts, and fine-tuning these prompts to optimize performance. The main emphasis is on developing more adaptable prompts. Fig 3.3 represents our proposed approach.

The text query and the dataset are sent to the RAG which contains a vector base and converts this dataset and query into their corresponding embeddings. These embeddings from RAG are then sent to the LLM Mistral. Prompts are designed by this dataset file, we have used few shot prompt engineering method, we take examples from the dataset file and give instructions to the mistral to provide these legal entities which are given in the example prompts. Now Mistral has embeddings as a context of the legal domain terminologies and expressions

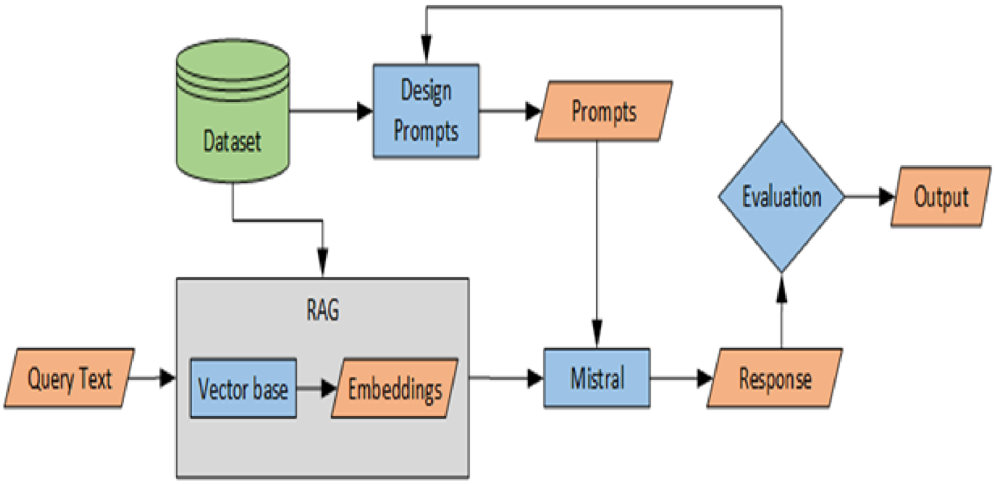


FIGURE 3.3: Diagram of Proposed System

and prompts and it provide the legal entities. These entities are then evaluated if their result is not satisfactory again prompts are designed and then these prompts are sent again to mistral, then the result is again evaluated, if it is satisfactory then this would be our final output. So, it is an iterative process which enhances model's performance.

3.2.1 Large Language Models

LLMs like GPT and BERT are powerful tools in NLP that excel in understanding and interpreting complex language structures. These models are trained on vast and varied datasets, which equips them with the ability to handle intricate language patterns often found in specialized fields such as legal, medical, or technical do- mains. This capability is crucial for accurate NER, where the goal is to identify and classify specific entities within text, such as legal terms, medical conditions, or technical jargon.

LLMs like GPT and BERT can be fine-tuned on domain-specific datasets to enhance their performance in these specialized contexts. Fine-tuning involves additional training of a pretrained model using a dataset that is representative of the target domain. For instance, to improve the model's ability to recognize legal entities, it would be trained on a dataset containing annotated legal texts with clearly defined entities. This process adjusts the model's parameters to better understand the context and specific language associated with the legal domain. The same approach applies to medical or technical domains, where fine-tuning the model on relevant datasets helps it learn to recognize domain-specific terminology and entities more effectively.

GEMMA is designed for generalized entity extraction and analysis, making it versatile for various NER applications, while Mistral is known for its capabilities in handling complex language tasks. In practice, the integration of LLMs with domain-specific fine-tuning allows for enhanced accuracy in NER by enabling the models to adapt to the unique linguistic characteristics and entity types present in specialized domains. This process typically involves using annotated datasets that include a diverse range of entity examples specific to the field, allowing the models to learn and generalize effectively. By leveraging these advanced models and techniques, researchers and practitioners can achieve higher precision and relevance in entity recognition tasks, supporting more effective analysis and interpretation of complex texts across various domains, also LLMs are adaptable and scalable models across various domains.

3.2.2 Retrieval Augmented Generation

RAG enhances NER by incorporating a retrieval mechanism that draws relevant information from a large corpus of documents to provide context for the generative model. In traditional NER tasks, models are often limited to the data they have been trained on, which can result in challenges when dealing with rare, ambiguous, or domain- specific entities. However, RAG's retrieval system compensates for this limitation by dynamically accessing external information, enriching the model's understanding of the entities it encounters.

Retrieval, the process of accessing and utilizing relevant information, has been demonstrated to enhance performance across a diverse range of NLP tasks. These tasks encompass open-domain question answering, fact checking, fact completion, long-form question answering, Named Entity Recognition, Relation Extraction, Knowledge Graph Construction, dialogue, translation, and language modeling [51].

For NER tasks in specialized fields like legal, medical, or technical domains, this retrieval component plays a critical role. Legal texts, for example, often contain complex entities such as case law references, statutes, or legal terms that can be highly ambiguous or context-dependent. A standalone model might struggle to disambiguate such entities, particularly if they have multiple meanings or are uncommon in the training data. By retrieving relevant legal cases, statutes, or other documents, RAG provides the necessary context to clarify these ambiguities. This allows the model to correctly interpret and classify entities based on the broader corpus of information, rather than relying solely on the input text.

In the legal domain, for instance, RAG could retrieve related cases or legal provisions that mention a specific statute or court ruling, which helps the model distinguish between different interpretations of an entity based on the context. This dynamic retrieval of supplemental data improves the accuracy and reliability of the NER model, as it can now make more informed predictions about entity types and relationships [23].

Moreover, RAG enhances the overall comprehension of text, as the retrieval component brings in relevant background knowledge that supports more coherent entity classification. This process is particularly important for NER tasks where entities are interconnected and depend heavily on external references or domain knowledge. By incorporating this information, RAG not only improves the precision of entity recognition but also fosters more meaningful applications in professional fields, such as legal research, medical documentation, and technical writing.

In educational settings, for example, NER tasks using RAG can assist students or researchers by automatically pulling in additional resources related to specific entities, helping them understand complex terminology or concepts more deeply. In professional environments, such as law firms or healthcare institutions, RAG-powered NER systems can streamline document analysis by ensuring that key entities are recognized with greater accuracy and context, reducing the need for manual review and interpretation [35]

Ultimately, the combination of retrieval-based augmentation and generative capabilities in RAG leads to more reliable, coherent, and impactful NER applications. By pulling in relevant information from large corpora in real-time, RAG enhances the performance of NER models, making them more adaptable and accurate across a wide range of specialized domains.

3.2.3 Prompt Engineering

In the context of LNER, carefully crafted prompts play a crucial role in directing LLMs like GPT or BERT to accurately identify specific types of legal entities. The legal domain is complex and often involves highly specialized terminology, including court names, case numbers, statutory references, legal citations, and other domain-specific entities. By employing prompt engineering, LLMs can be guided to focus on extracting these entities with higher precision.

For instance, when processing legal documents, a prompt can be designed to instruct the model to prioritize the identification of entities related to judicial proceedings. This could include court names, such as" Supreme Court of the United States," case numbers like" No. 18-1234," or references to statutes such as" Section 101 of the Patent Act." The model, when prompted with such instructions, can more effectively zero in on the legal entities that are most relevant to the task at hand.

This becomes particularly useful when parsing through extensive legal texts where the distinction between different types of entities is crucial for accurate legal analysis or document summarization.

Prompt engineering for LNER can also include more sophisticated instructions to capture the unique context in which legal entities appear. For example, certain prompts could instruct the model to focus on identifying entities that appear in specific sections of legal documents, such as those dealing with court rulings, case law, or legal precedents.

Additionally, prompts can be tailored to recognize variations in legal terminology, which can be challenging for models not explicitly trained on legal datasets. For example, legal citations can take various forms depending on jurisdiction, case type, or document format. A prompt could instruct the model to recognize and extract these variations, whether it's a formal citation like" Brown v. Board of Education, 347 U.S. 483 (1954)" or an informal reference embedded within legal arguments. By specifying these types of entities in prompts, the model becomes more adept at identifying even subtle differences in legal entity formatting or nomenclature.

Moreover, prompt engineering offers a level of flexibility that can adapt to different legal subdomains [34] For example, a prompt might be fine-tuned to focus on corporate law, directing the model to identify entities such as company names, mergers, or legal contracts. Alternatively, in criminal law contexts, prompts could guide the model to focus on criminal statutes, defendant names, or sentencing information. This adaptability allows LNER systems to perform more accurately across various legal sectors, ensuring the correct identification of entities that are specific to each subdomain.

By carefully crafting these prompts, NER models become more targeted in their extraction tasks, enhancing their ability to recognize legal entities with greater accuracy. This approach reduces the noise that can occur when models process general or ambiguous data, as prompts provide the guidance necessary to narrow the scope of analysis. As a result, prompt engineering plays a pivotal role in refining NER outputs, making them more reliable for legal professionals, researchers, and institutions that rely on accurate legal entity recognition for tasks like document drafting, legal research, and case analysis.

Prompt 1

Here I have generally asked the model to extract legal entities from the text and it extracts some generic entities like person, place, organization and date. As shown in figure 3.4

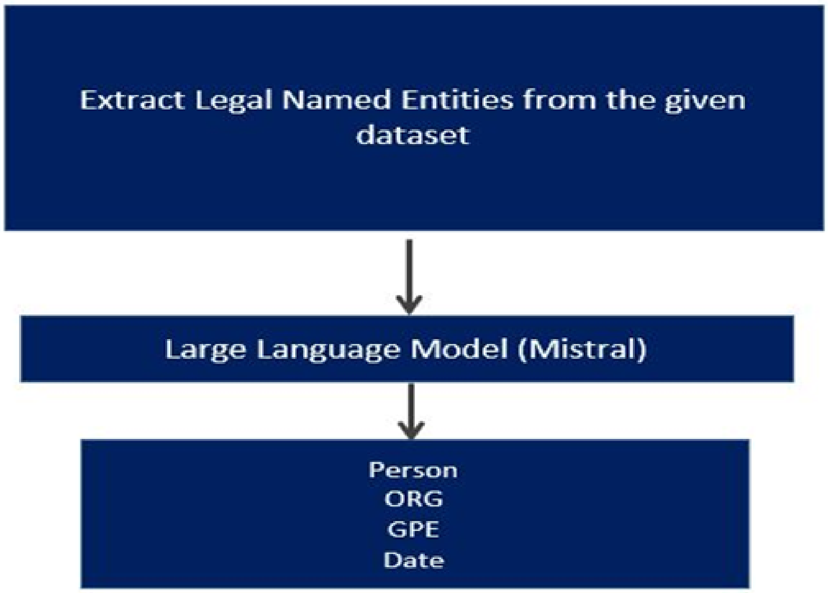


Figure 3.4: Prompt 1

Prompt 2

Here in Figure 3.5, I have given an example of a text which contains legal entities and I have guided the model that these are the 14 entities to find in the text.

Model will have better understand that which person is lawyer, judge, respondent or witness etc.

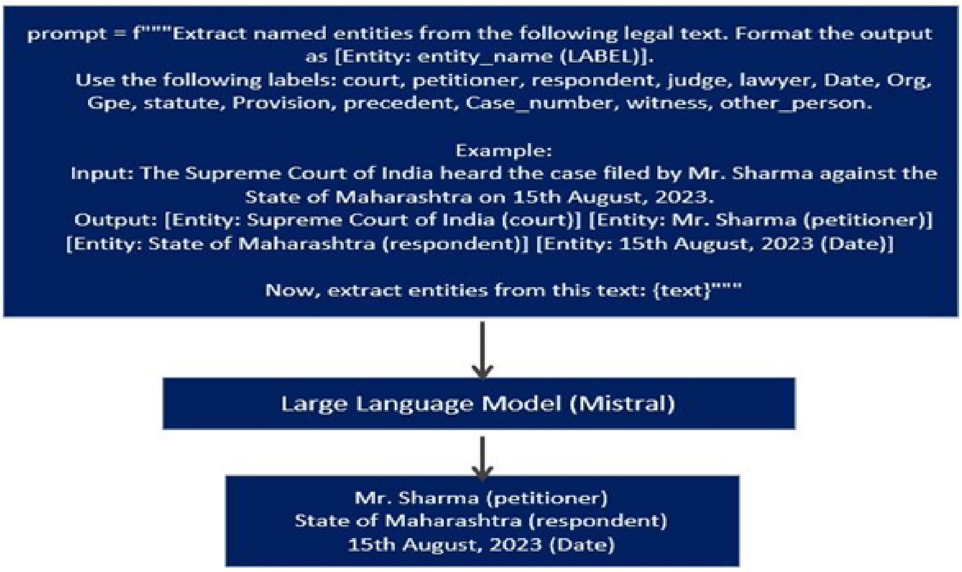


Figure 3.5: Prompt 2

Prompt 3

For prediction, prompt 3 contains the example of two entities that both are ORG. Here is the pictorial representation of prompt 3 in figure 3.6.

The text in dataset is "On specific query by the Bench about an entry of Rs. 1,31,37,500 on deposit side of Hongkong Bank account of which a photo copy is appearing at p. 40 of assesses paper book, learned authorized representative submitted that it was related to loan from broker, Rahul Co. on the basis of his submission a necessary mark is put by us on that photo."

Then prompt was designed to instruct the model to extract entities.

In this sentence both the entities are Organization that are Hongkong Bank and Rahul Co.

So in prompt we have instruct the model by providing label along with these entities. So that model understand that it has to extract these entities from the dataset with these labels.

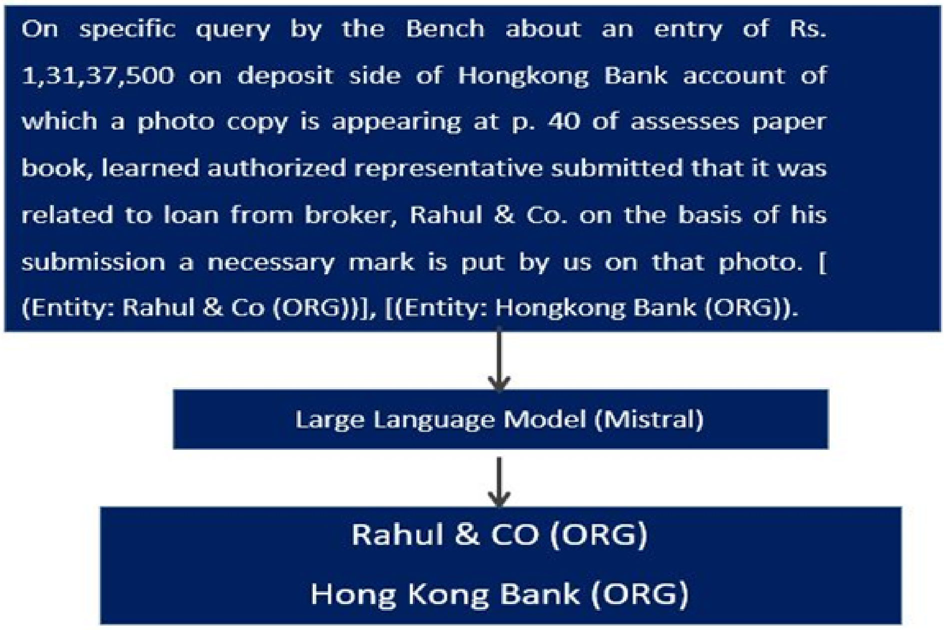


Figure 3.6: Prompt 3

Prompt 4

For prediction, prompt 4 contains the example of two entities that both are Other_Person. Here is the pictorial representation of prompt 4 in Figure 3.7. The text is dataset is "He was also asked whether Agya Kaur, mother-in-law of the deceased lived separately from Tarlochan Singh." and then prompt was designed.

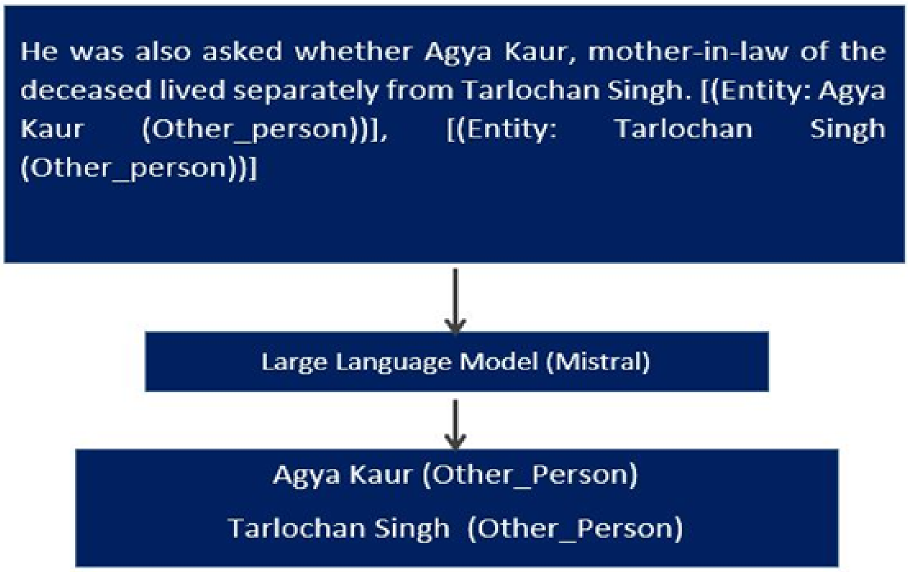


FIGURE 3.7: Prompt 4

Prompt 5

This prompt contains example of 4 entities, two of which are ORG, one is WITNESS and the other is GPE.The text in dataset is "Mr Vijay Mishra, Deputy Manager, HDFC Bank, Noida, UP has deposed that complainant had a current account with HDFC Bank in the year 2004." then prompt was made.

Here is the pictorial representation of prompt 5 in Figure 3.8.

This example of prompt is instructing the model about Organization and witness entities, this example is taken from training dataset.

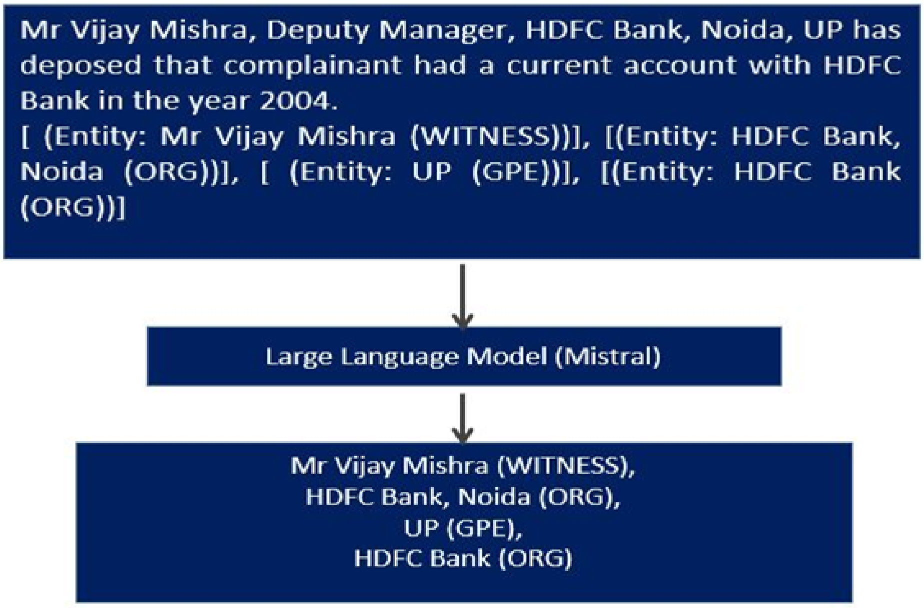


FIGURE 3.8: Prompt 5

Prompt 6

Here is the pictorial representation of prompt 6 in Figure 3.9. The actual text is "If the argument of the learned counsel for the respondents is accepted, it would mean that a person whose bail under POTA has been rejected by the Special Court will have two remedies and he can avail any one of them at his sweet will."

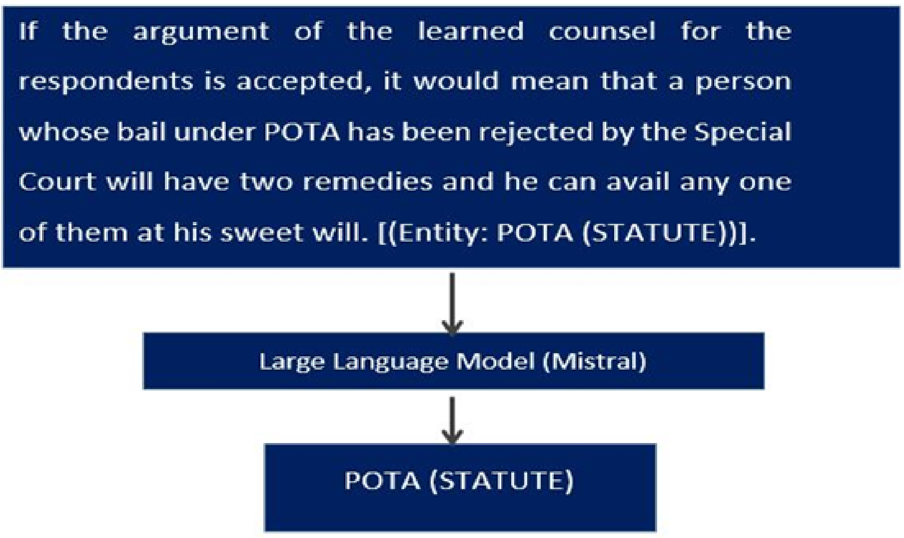


FIGURE 3.9: Prompt 6

Prompt 7

This prompt contains examples of the entities court and date. Here is the pictorial representation of prompt 7 in Figure 3.10. The original test from dataset is " In view of these fact, it can he safely held that for the area of Jaipur City (municipal limits) there were two Special Judges on 29.4.1968, one was Sessions Judge Jaipur City, who by virtue of his Office was appointed to be Special Judge for that area by virtue of-Notified ion dated 26.2.1968." Then the prompt has been designed.

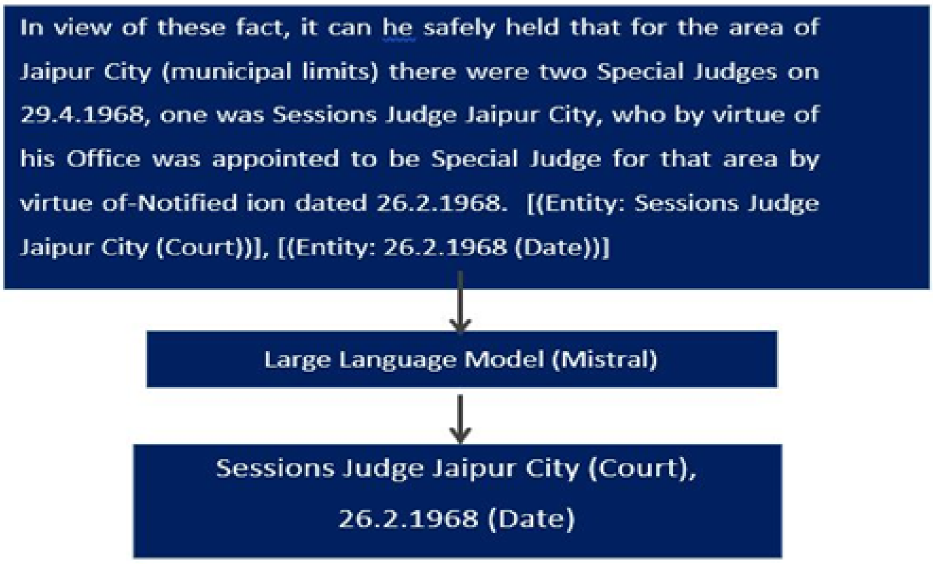


FIGURE 3.10: Prompt 7

Prompt 8

This prompt contains examples of the entities Court and Precedent. Here is the pictorial representation of prompt 8 in Figure 3.11.

This example of prompt is instructing the model about Court and Precedent entities, this example is taken from training dataset.

From result We have examined that with this prompt Models was predicting most of the entities with label Court and Precedent correctly.

The original text in dataset is "The Supreme Court, in the case of Susamma Thomas, 1994 ACJ 1 (SC), has awarded a sum of Rs. 15,000/- each on the above two heads."

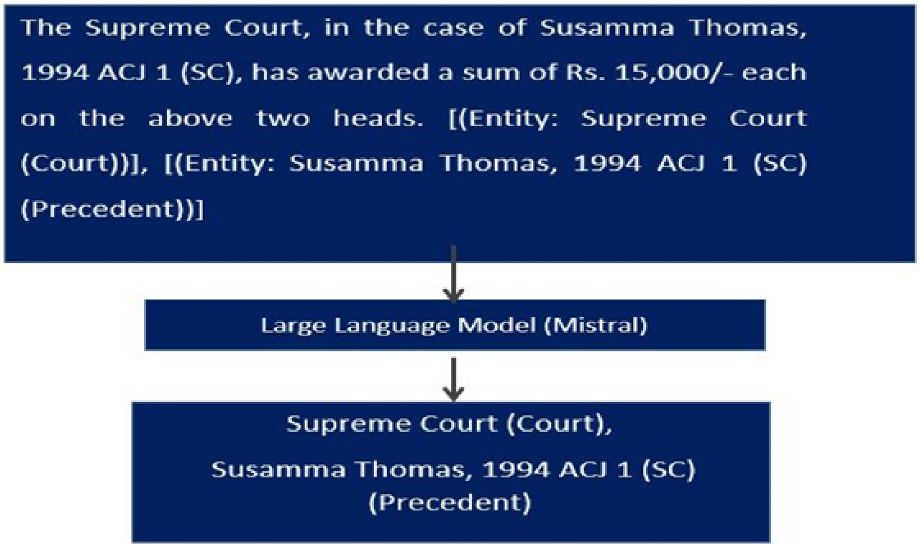


FIGURE 3.11: Prompt 8

Prompt 9

This prompt contains examples of the entities Respondent, GPE, Date. Here is the pictorial representation of prompt 9 in Figure 3.12.

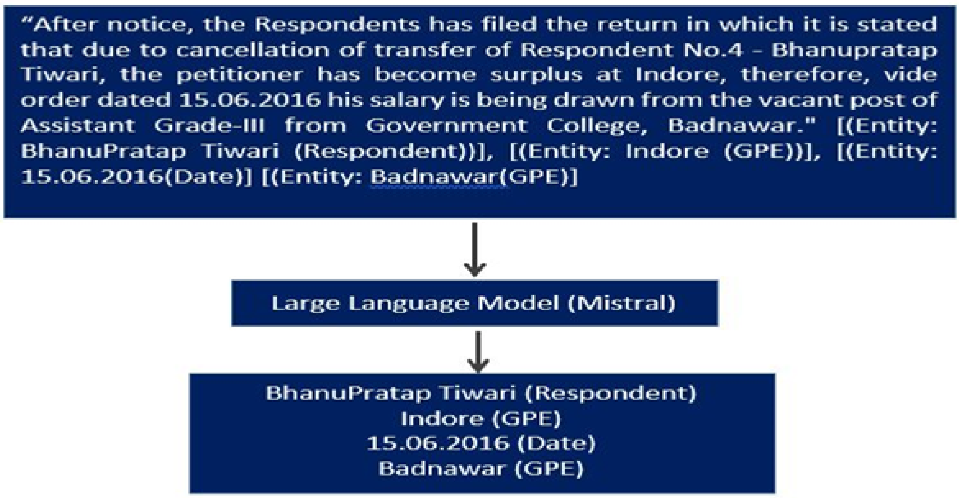


FIGURE 3.12: Prompt 9

Prompt 10

This prompt contains examples of the entities Provision and GPE. Here is the pictorial representation of prompt 10 in Figure 3.13.

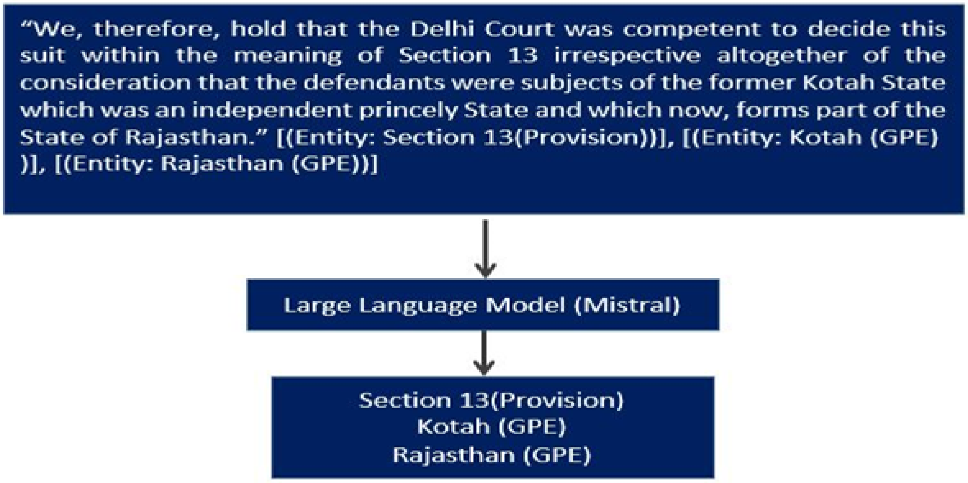


FIGURE 3.13: Prompt 10

Prompt 11

This prompt contains examples of the entities Court, date and judge. Here is the pictorial representation of prompt 11 in Figure 3.14.

This example of prompt is instructing the model about Court, date and Judge entities, this example is taken from training dataset.

From result We have examined that with this prompt Models was predicting most of the entities with label Court, date and Judge correctly.

So careful prompts enhance model's performance.

Model will better understand that which are the court names and what different types of being used in the court ruling and what are the judge names extracted from both preamble and judgment section of dataset.

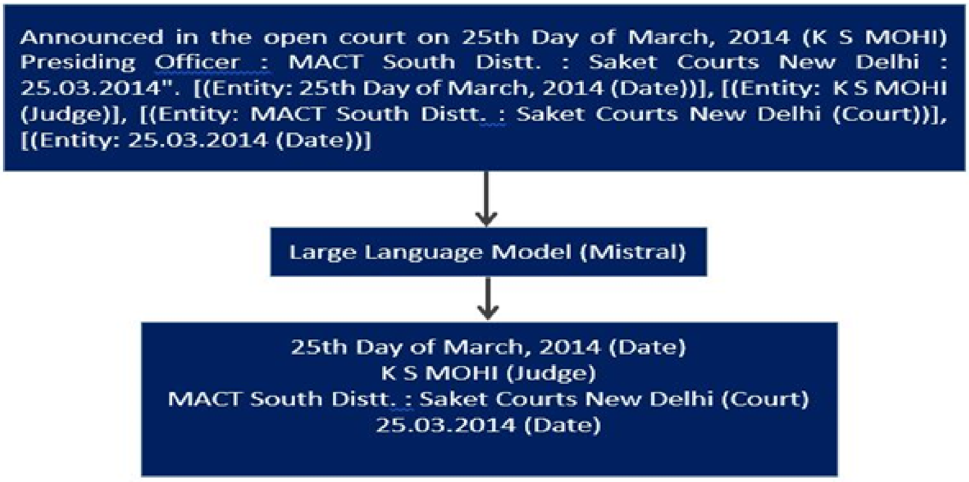


FIGURE 3.14: Prompt 11

Prompt 12

This prompt contains examples of the entities Org Other_person. Here is the pictorial representation of prompt 12 in Figure 3.15. This example is taken from training dataset.

From result We have examined that with this prompt Models was predicting most of the entities with label Org and Other_person correctly.

So careful prompts enhance model's performance.

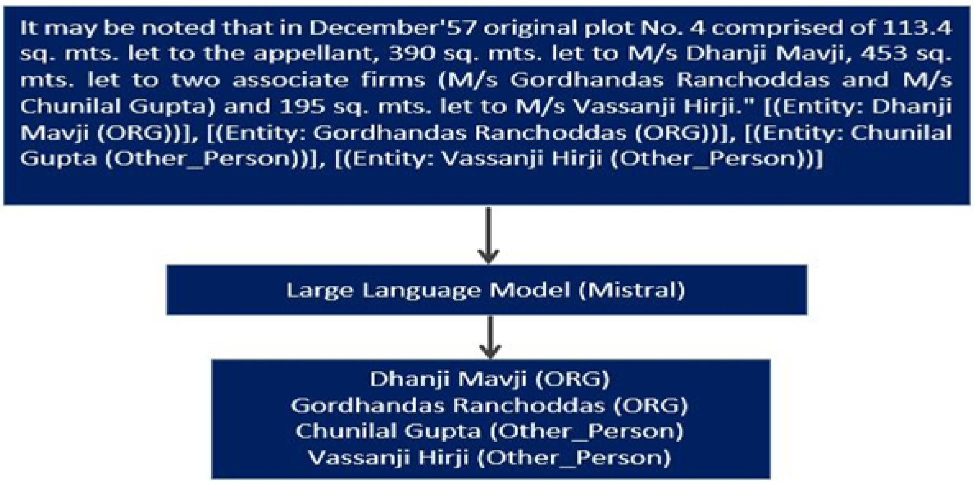


FIGURE 3.15: Prompt 12

Chapter 4

Experimentation

The experiments for this study were conducted using free version of Kaggle, a cloud-based platform that provides a convenient and accessible environment for data scientists and machine learning practitioners under Google LLC. Kaggle was using Python 3.10.12 at the time of experimentation. Kaggle typically provides around 29 GB of RAM and access to a standard CPU, also provides different types of accelerators. I used GPU T 100 for this experiment. For storage, users have approximately 57 GB of local temporary disk space.

The libraries and packages necessary for implementing this code are as follows:

- transformers
- langchain-community
- langchain-core
- import HuggingFacePipeline
- import AutoTokenizer
- import HuggingFaceEmbeddings
- import FAISS
- import RetrievalQA

- from langchain.prompts import PromptTemplate
- import PyPDFDirectoryLoader
- import RecursiveCharacterTextSplitter
- import textwrap
- import login
- import JSONLoader
- import SentenceTransformer
- import faiss
- import pandas as pd
- import torch
- import re
- import tqdm
- from sklearn.metrics import f1_score
- import csv
- import osjson

In the experiment these libraries and packages were used. Mistral Model was used from hugging face platform and RAG is used using langchain components. Langchain and hugging face community supports different NLP tasks effectively. It contains all required libraries, dictionaries and packages.

We have started to work with the model "en_legal_ner_trf" which is a transformer model. The dataset InLegalBert is trained over this model.

Following are some of the results of judgements using "en_legal_ner_trf".

Precision = 0.8826

Recall = 0.9019

F1 Score = 0.8921

The model is trained upon the same data, but problem is when the dataset is changed the model can not perform well over it.

We need to annotate the incoming data, we need to annotate data when it contain more number of entities rather these fourteen entities used above. We may not be able to perform other NLP tasks like question answering, text summarization etc, as model performance significantly decreases if it is not trained for other tasks.

So the solution is to use LLM which is trained over billion of parameters and adapt themselves according to evolutions occurring in the day by day languages and in terminologies of legal domain.

We used RAG which will provide a context to the LLM Mistral for the efficient named entity recognition task. Comprehensive prompts that includes examples from the dataset will instruct the model to extract those fourteen number of entities mentioned above.

Here the dataset is in json format, but simple word, text or pdf document can also be passed to the model to provide it a contest. So NER with LLM is not dataset dependent. Moreover same document can also be used to perform other tasks like relationship extraction, knowledge graph construction, question answering systems, text summarization etc.

This technique is also scalable to the other subdomains of legal domain. Same method of NER can also be used for criminal, cyber forensics , contract documents and other court rulings. We have to make prompts according to the dataset. We can include some examples of the entities to enhance model's performance for NER.

4.1 Evaluation

4.1.1 Evaluation Measure

The performance of NER systems is assessed using precision, recall, and the F1 measure. The F1 score, also known as the F measure, represents the harmonic mean of precision and recall.

$$\text{F1 Score} = 2.\frac{Precision.Recall}{Precision + Recall}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the system. It indicates how reliable the system's positive predictions are. Whereas, Recall measures the proportion of true positive predictions made by the system out of all the actual positive instances. It indicates how well the system is able to identify positive cases.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

where:

- TP = True Positives
- TN = True Negative
- FP = False Positives
- FN = False Negatives

4.2 Results

Table 4.1 shows that proposed system has achieved an F1 score to 0.68 along with Large Language Model Mistral and Prompt Engineering system.

Sr. No.	**Author**	**Method**	**F1 Score**
1	Proposed System	LLMs+Prompts+RAG	0.68
2	Atin et al. [6]	Mistral	0.63

Table 4.1: Results

The improvement in results, reflected by the increase in the F1 score from 0.63 to 0.68 after applying carefully crafted prompts and integrating RAG, underscores significant advancements in the performance of LNER systems. The results, we have achieved using RAG, Mistral and prompt engineering methods indicates that they can also serve in the field of information retrieval without needing the domain specific datasets like the previous ML and Rule-based systems needed. Moreover they are adaptable to the other subdomains and are scalable to the other NLP tasks. Because these models are adaptable to the evolving terminologies and transformations.

Chapter 5

Conclusion and Future Work

In this thesis, an extensive review of existing research on subdomain of NLP called NER systems has been conducted, offering critical insights into the strengths and limitations of various approaches. The literature reveals that much of the current research heavily relies on machine learning and deep learning techniques, which, while powerful, require vast amounts of training data and considerable computational resources and also these techniques are not scalable and adaptable to the subdomains of legal domain. These techniques are dataset or format dependent, if dataset is not in proper format these techniques may not perform well for the task ahead. So we used a technique which is not dataset dependent.

A notable advancement presented in this thesis is the use of a hybrid system that integrates prompt engineering with the RAG approach. A large language model Mistral is used for the legal named entity recognition task and it provides F1 score of 0.68 which is better than the other techniques that have been previously used.

Future work in this area is to perform named entity recognition task for Pakistani court data and also work on other legal entities rather than these fourteen entities of the dataset used. We also want to work on relationship extraction, then populate a knowledge graph and develop a chat bot which will be capable to answer the questions of legal domain and which may be able to provide a voice interface too.

Bibliography

[1] Satoshi Sekine. Named entity: History and future. *Project notes, New York University*, 4, 2004.

[2] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2(e93), 2016.

[3] Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Elena Baralis, Luca Cagliero, and Francesco Tarasconi. Politohfi at semeval2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1401–1411, Toronto, Canada, 2023. Association for Computational Linguistics.

[4] Archana Goyal, Vishal Gupta, and Manish Kumar. Deep learning-based named entity recognition system using hybrid embedding. *Cybernetics and Systems*, 55(2):279–301, 2024.

[5] A. Thomas and S. Sangeetha. Knowledge graph based question-answering system for effective case law analysis. In *Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, pages 291–300. Computer Linguistics, 2022.

[6] Atin Sakkeer Hussain and Anu Thomas. Large language models for judicial entity extraction: A comparative study, 2024.

[7] Davlatyor B Mengliev, Vladimir B Barakhnin, Mukhammadjon Atakhanov, Bahodir B Ibragimov, Mukhriddin Eshkulov, and Bobur Saidov. Developing rule-based and gazetteer lists for named entity recognition in uzbek language: geographical names. In *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*, pages 1500–1504. IEEE, 2023.

[8] J Ravikumar and Kumar P Ramakanth. Machine learning model for clinical named entity recognition. *International Journal of Electrical and Computer Engineering*, 11(2):1689, 2021.

[9] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. A knowledge-enhanced pretraining model for legal ner. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3347–3358. International Committee on Computational Linguistics, 2020.

[10] Duraimurugan Rajamanickam. Improving legal entity recognition using a hybrid transformer model and semantic filtering approach. *arXiv preprint arXiv:2410.08521*, 2024.

[11] Anu Thomas and S Sangeetha. Deep learning architectures for named entity recognition: A survey. In *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1*, pages 215–225. Springer, 2020.

[12] Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. A survey on recent advances in sequence labeling from deep learning models. *arXiv preprint arXiv:2011.06727*, 2020.

[13] Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. A novel prompting method for few-shot ner via llms. *Natural Language Processing Journal*, 8:100099, 2024.

[14] Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. SemEval-2023 task 6: LegalEval - understanding legal texts.

In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2362–2374, Toronto, Canada, July 2023. Association for Computational Linguistics.

[15] Dhanachandra Ningthoujam, Pinal Patel, Rajkamal Kareddula, and Ramanand Vangipuram. Researchteam_hcn at semeval-2023 task 6: A knowledge enhanced transformers based legal nlp system. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1245–1253. Association for Computational Linguistics, 2023.

[16] Quang-Minh Tran and Xuan-Dung Doan. Vtcc ner at semeval-2023 task 6: An ensemble pre-trained language models for named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 415–419, Toronto, Canada, 2023. Association for Computational Linguistics.

[17] Intisar Almuslim, Sean Stilwell, Surya Kiran Suresh, and Diana Inkpen. uottawa at semeval-2023 task 6: Deep learning for legal text understanding. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada, 2023. Association for Computational Linguistics.

[18] Yuri Noviello, Enrico Pallotta, Flavio Pinzarrone, and Giuseppe Tanzi. Teamunibo at semeval-2023 task 6: A transformer-based approach to rhetorical roles prediction and ner in legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 275–284, Toronto, Canada, 2023. Association for Computational Linguistics.

[19] Jingjing Huo, Kezun Zhang, Zhengyong Liu, Xuan Lin, Wenqiang Xu, Maozong Zheng, Zhaoguo Wang, and Song Li. Antcontenttech at semeval-2023 task 6: Domain-adaptive pretraining and auxiliary task learning for understanding indian legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 402–408, Toronto, Canada, 2023. Association for Computational Linguistics.

[20] Michael Ginn and Roman Khamov. Ginn khamov at semeval-2023 task 6, subtask b: Legal named entities extraction for heterogenous documents. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 420–425, Toronto, Canada, 2023. Association for Computational Linguistics.

[21] Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. Named entity recognition in Estonian 19th century parish court records. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313, Marseille, France, June 2022. European Language Resources Association.

[22] N. Ahmed, S. Latif, R. Irfan, A. Ul-Hasan, and F. Shafait. Comparison of transformer models for information extraction from court room records in pakistan. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 01–06. IEEE, 2022.

[23] B. Sarmah, B. Hall, R. Rao, S. Patel, S. Pasquali, and D. Mehta. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction, 2024.

[24] Yucheng Hu and Yuxing Lu. HW-TSC at SemEval-2023 task 7: Exploring the natural language inference capabilities of ChatGPT and pre-trained language model for clinical trial. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1603–1608, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25] M. Soto-Jiménez, J. Martínez-Velásquez, P. Chicaiza, P. Vinueza-Naranjo, and N. Bouayad-Agha. Rag-based question-answering systems for closed-domains: Development of a prototype for the pollution domain. In *Intelligent Systems Conference*, pages 573–589. Springer, 2024.

[26] J. He, S. Tallam, S. Nakshathri, N. Amarnath, P. Kr, and D. Kumar. Infrrd.ai at semeval-2024 task 7: Rag-based end-to-end training to generate headlines and numbers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 940–951, 2024.

[27] Shiye Li and Li Yi. A few-shot entity relation extraction method in the legal domain based on large language models. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (DEAI '24)*, pages 580–586, New York, NY, USA, 2024. Association for Computing Machinery.

[28] Gustavo Henrique de Araújo, Thiago de Campos, and Fernando de Assis Braz. Information extraction from legal documents using linguistic knowledge and ontologies. In *Proceedings of the 13th International Conference on Information Systems and Technology (ICIST)*. SciTePress, 2021.

[29] Fabián Villena, Luis Miranda, and Claudio Aracena. llmner: (zero—few)-shot named entity recognition, exploiting the power of large language models, June 2024.

[30] H. Bonifacio, P. A. Vilela, G. R. Lobato, and E. R. Fernandes. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 648–662. Springer, 2020.

[31] T. et al Brown. Language models are few-shot learners. In H. Larochelle, Ranzato M., Hadsell R., M.F. Balcan, and H. (eds.) Lin, editors, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Curran Associates, 2020.

[32] P. Liu, W. Yuan, and J. et al. Fu. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, volume 55, pages 1–35, 2023.

[33] Serena Villata, Michal Araszkiewicz, Kevin Ashley, Trevor Bench-Capon, L Karl Branting, Jack G Conrad, and Adam Wyner. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, pages 1–31, 2022.

[34] Y. Yang, Z. Wang, and Z. Jiang. An improved tri-training based named entity identification approach for legal knowledgebase of properties involved in criminal cases. In *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, pages 655–660. IEEE, 2021.

[35] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, September 2024.

[36] Luis Adrián Cabrera-Diego and Akshita Gheewala. Jus mundi at semeval-2023 task 6: Using a frustratingly easy domain adaption for a legal named entity recognition system. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada, 2023. Association for Computational Linguistics.

[37] C. Çetindağ, B. Yazıcıoğlu, and A. Koç. Named-entity recognition in turkish legal texts. *Natural Language Engineering*, 29(3):615–642, 2023.

[38] Pin Tang, Yi Zhou, Feng Lin, Yan Wang, and Yanwei Zhang. Recognizing chinese judicial named entity using bilstm-crf, 2020.

[39] H. Keshavarz et al. Named entity recognition in long documents: An end-to-end case study in the legal domain. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2024–2033, Osaka, Japan, 2022. IEEE.

[40] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[41] Junzhe Zhao, Yingxi Wang, Nicolay Rusnachenko, and Huizhi Liang. Legal_try at semeval-2023 task 6: Voting heterogeneous models for entities identification in legal documents. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1282–1286. Association for Computational Linguistics, 2023.

[42] Sébastien Bosch, Laurent Estève, José Loo, and Anne-Laure Minard. Uoloutal at semeval-2023 task 6: Lightweight systems for legal processing. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 489–496. Association for Computational Linguistics, 2023.

[43] Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 172–179. ACL Anthology, 2021.

[44] Harshil D. Joshi, Deepak G. Gohil, Abhishek R. Kher, and Pritesh P. Patel. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges, 2022.

[45] E. Leitner, G. Rehm, and J. Moreno-Schneider. Fine-grained named entity recognition in legal documents. In *International conference on semantic systems*, pages 272–287. Springer, 2019.

[46] Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, 2020.

[47] Raabia Mumtaz and Muhammad Abdul Qadir. Custner: A rule-based named-entity recognizer with improved recall. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 16(3):110–127, 2020.

[48] Xiang Jin and Yingqi Wang. Teamshakespeare at semeval-2023 task 6: Understand legal documents with contextualized large language models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 517–525, Toronto, Canada, 2023. Association for Computational Linguistics.

[49] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, and Quentin Malartic. The falcon series of open language models, 2023.

[50] V. Oliveira, G. Nogueira, and T. et al. Faleiros. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, 2024.

[51] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Appendix

Some examples of Court Judgments

Example 1: Indian Supreme Court Judgment Sample

Lifeforce Cryobank Sciences Inc Vs. Cryoviva Biotech Pvt. Ltd. & Ors.

[**Arbitration Petition No. 15/2018**]

Manoj Misra, J.

1. The petitioner which is a company duly incorporated under the laws of the United States of America1 has invoked the jurisdiction of this Court under sub-sections (6) and (12) of Section 11 of the Arbitration and Conciliation Act, 19962 for appointment of a sole arbitrator in terms of arbitration clause stipulated in the agreements dated 27 December 2009 and 11 February 2010 to adjudicate upon the disputes between the petitioner and the respondents.

2. The petitioner's case inter alia is that it has purchased the assets of Cryobank International, Inc3 on 8 June 2010 at a public auction in pursuance of a decree dated 5 May 2010 passed by the Circuit Court of Florida, USA. Following which, a certificate of title was issued in its favor certifying purchase of all assets, tangible and intangible, of Cryobank USA by it. On basis thereof, the petitioner claims to have stepped into the shoes of Cryobank USA.

3. According to the petitioner, the dispute between the petitioner and the respondents stems from Exclusive and Perpetual License Agreement4 and Share Subscription and Shareholders Agreement.5 License agreement is between Cryobanks USA and Cryobanks India International Pvt. Ltd (now known as Cryoviva Biotech Pvt. Ltd. - Respondent No.1 herein). The same contains an arbitration clause in Section 7.

Whereas Share Subscription Agreement is between RJ Corp (respondent no.2 herein) acting on behalf of itself and its shareholders, namely, Devyani Enterprises Pvt. Ltd.- respondent no.3, Devyani Overseas Private Ltd. - respondent no.4, RK Jaipuria & Sons (HUF) - respondent no.5, Dhara Jaipuria - respondent no.6; Cryobank USA; and Cryobanks India International Pvt Ltd (now Cryoviva Biotech Pvt. Ltd. - Respondent No.1). The same has an arbitration clause in clause XVII. Under both the arbitration agreements the disputes are referable to a sole arbitrator subject to the jurisdiction of courts at Delhi.

4. It is the case of the petitioner that under the license agreement, the respondents were entitled to use Cryobank's intellectual property rights in lieu of consideration which included issue of shares in the respondent company. It is stated that the petitioner stepped into the shoes of Cryobank USA, and this fact was acknowledged by the respondent company in various correspondences. However, since petitioner's demand was not met, arbitration clause had to be invoked vide notice dated 29.09.2017.

5. In response to the notice of these proceedings, the respondents' case inter alia is that the license agreement was non-assignable, and the respondents have not accepted the petitioner as the assignee. There is, therefore, no privity of contract. Hence, the petition is liable to be dismissed.

6. We have heard learned counsel for the parties.

7. At the stage of considering an application for appointment of an arbitrator the Court is required to examine whether there exists an arbitration agreement between the parties. The existence of an arbitration agreement is not an issue.

The issue is that it is not between the petitioner and the respondent company but between Cryobank USA and the respondents. According to the respondents the petitioner has only bought assets of Cryobank USA but, in absence of respondents' consent, has not stepped into the shoes of Cryobank USA.

8. On the other hand, the petitioner has referred to several documents/correspondences to canvass that the respondent has accepted the petitioner as having stepped into the shoes of Cryobank USA. Petitioner has also annexed certificate to indicate that rights under all existing contracts including intellectual property rights of Cryobank USA were purchased by the petitioner in auction sale.

9. In Khardah Company Ltd. v. Raymon & Co (India) Pvt. Ltd., AIR 1962 SC 1810 it was held that an assignment of a contract might result by transfer either of the rights or of the obligations thereunder. But there is a well-recognized distinction between these two classes of assignments. As a rule, obligations under a contract cannot be assigned except with the consent of the promisee, and when such consent is given, it is really a novation resulting in substitution of liabilities. On the other hand, the rights under a contract are assignable unless the contract is personal in its nature, or the rights are incapable of assignment either under the law or under an agreement between the parties.

10. Following the decision in Khardah Company (supra) in DLF Power Ltd. v. Mangalore Refinery & Petrochemicals Ltd., 2016 SCC OnLine Bom 5069 a single judge of the Bombay High Court held that the arbitration agreement in a contract is a benefit which can be assigned along with the main contract or even otherwise.

11. Be that as it may, since at the stage of consideration of a prayer under Section 11(6) of the 1996 Act the Court has to confine itself to the examination of the existence of an arbitration agreement (vide sub-section (6-A) of Section 11), it would not be appropriate for us to delve deep into the issue as it could well be considered by the arbitrator on the basis of evidence led by the parties. More so, when existence of arbitration agreement in the license agreement and share subscription agreement is not in dispute.

12. We, therefore, deem it appropriate to refer the matter to the Delhi International Arbitration Centre for appointment of a sole arbitrator to adjudicate upon the dispute between the parties.

13. It is made clear that we have not expressed any opinion on the merits of the claim of either party including with regard to the arbitrability of the dispute. All contentions and pleas are kept open for the parties to raise before the arbitral tribunal.

14. Subject to above, the petition including all pending applications, if any, stand disposed of.

.....................CJI. (Dr. D.Y. Chandrachud)

.......................J. (Manoj Misra)

New Delhi;

November 8, 2024

Example 2: Pakistan Supreme Court Judgment Sample

IN THE SUPREME COURT OF PAKISTAN
(Review Jurisdiction)

Present:
Justice Qazi Faez Isa, CJ
Justice Muhammad Ali Mazhar
Justice Musarrat Hilali

Civil Review Petition No. 14 of 2024
In
Civil Petition No. 42 of 2024

Pakistan Tehreek-i-Insaf, through its authorized person and others. … *Petitioners*

Versus

Election Commission of Pakistan, through Special Secretary, Islamabad and others. … *Respondents*

For the Petitioners:	Nemo.
For Respondent No. 3:	Syed Ahmed Hassan Shah, ASC. a/w respondent No. 3.
Date of Hearing:	11.10.2024.

ORDER

Review of the short order of this Court, announced on 13 January 2024, and of its detailed reasons, which were issued on 25 January 2024 is sought through this review petition.

2. An '*Application for Adjournment*' has been submitted by Mr. Anis Muhammad Shahzad, Advocate-on-Record on the ground that, '*Mr. Hamid Khan, Senior ASC, has pressing family engagement at Lahore on 11.10.2024*'. The nature of the *pressing family engagement* is not disclosed. The learned AOR is also not in attendance. Learned counsel must know that merely filing an adjournment application does not mean that the case will be adjourned.

3. The petitioners were represented by Advocates of the Supreme Court ('**ASCs**') a couple of whom themselves are ASCs, respectively, Senior Advocate Mr. Hamid Khan, and ASCs Messrs Syed Ali Zafar, Gohar Ali Khan, Ajmal Ghaffar Toor, Niazullah Khan Niazi. Mr. Muhammad Sharif Janjua was the Advocate-on-Record earlier and in the review petition it is Mr. Anis Muhammad Shahzad. Therefore, if Senior Advocate Mr. Hamid Khan, as stated, had some *pressing family engagement* any of the other learned ASCs could have attended and proceeded with the case. Moreover, section 6 of the Supreme Court (Practice and Procedure) Act, 2023 now enables engagement of other counsel in a review petition, which was not permissible earlier.

4. In the circumstances, we are not persuaded to adjourn the case, however, in the interest of justice and only by way of indulgence we do so but make it clear that no further request for adjournment will be entertained, and we expect that the case to proceed on the next date.

5. In view of the fact that two Members of this Bench will not be available at Islamabad next week, we adjourn this case to Monday, 21 October 2024. Copy of this order be sent to the learned ASCs, learned AORs and to the review petitioners.

Chief Justice

Judge

Judge

Islamabad:
11.10.2024
(M. Tauseef)

Approved for reporting