A Comparative Study of Large Language Models for Named Entity Recognition in the Legal Domain

Tobias Deußer*†, Cong Zhao*†, Lorenz Sparrenberg*, Daniel Uedelhoven†, Armin Berger*†,
Maren Pielka*†, Lars Hillebrand†, Christian Bauckhage*†, Rafet Sifa*†
* University of Bonn, Bonn, Germany
† Fraunhofer IAIS, Sankt Augustin, Germany
tdeusser@uni-bonn.de
ORCID iD: 0000-0003-4685-0847

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825695

***Abstract*—Named Entity Recognition (NER) in the legal domain presents unique challenges due to specialized terminology and complex linguistic structures inherent in legal texts. While large language models (LLMs) like GPT-4, Llama-3, and others have significantly advanced natural language processing, their effectiveness in domain-specific tasks like legal Named Entity Recognition remains underexplored. This study conducts a comprehensive comparative analysis of eleven state-of-the-art LLMs on legal NER tasks across seven diverse datasets in five languages, namely English, Portuguese, German, Turkish, and Ukrainian. We evaluate the models' performance using F$_1$ scores, focusing on their ability to accurately identify and classify legal entities. Our findings reveal significant variability in LLM performance across different languages and legal contexts, with proprietary models like GPT-4 achieving the highest overall scores. The results highlight the influence of model architecture, dataset characteristics, and prompt design on the effectiveness of legal NER tasks. This study provides valuable benchmarks for legal NER applications and offers insights into the strengths and limitations of current LLMs, guiding future research and development in legal natural language processing.**

***Index Terms*—named entity recognition, large language models, legal domain, natural language processing, machine learning**

I. Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves identifying and classifying entities within text into predefined categories such as persons, organizations, locations, and legal statutes. Accurate NER is particularly crucial in the legal domain due to the specialized terminology and complex language structures inherent in legal documents. Effective NER facilitates information retrieval, legal analysis, and decision-making processes by extracting pertinent information from vast amounts of legal text.

The advent of large language models (LLMs) like GPT-4 [1], Llama-3 [2], or Phi-3 [3] has significantly advanced the field of NLP. These models have demonstrated remarkable performance across various tasks like sentiment analysis [4]–[6], contradiction detection [7]–[9], hate-speech detection [10], recommender systems [11], and, of course, general language understanding and generation [12] by leveraging deep learning techniques and extensive training data. However, applying LLMs to domain-specific tasks like legal NER presents unique challenges. Legal texts often contain archaic language, domain-specific jargon, and intricate syntactic constructions that may not be adequately captured by models trained on general language corpora.

This study aims to fill this gap by conducting a comprehensive comparative analysis of various state-of-the-art LLMs for NER in the legal domain. We evaluate general-purpose models to assess their effectiveness in recognizing and classifying legal entities. Our research investigates the following questions:

1) How do general-purpose LLMs perform on legal NER tasks?
2) How do different languages affect the performance of multilingual large language models?
3) Which model architectures are most effective for capturing the nuances of legal language in NER tasks?

By addressing these questions, we aim to provide insights into the strengths and limitations of current LLMs in legal NER applications. Specifically, we focus on evaluating eleven state-of-the-art LLMs across multiple datasets, including three English datasets and four datasets in other languages, to establish comprehensive benchmarks for legal NER tasks. The findings of this study are expected to guide future research and development efforts in legal NLP, assisting practitioners in selecting and optimizing models for enhanced performance in legal text analysis.

In the following, we discuss related works. Then, Section III describes our methodology. We shed light on our experiments and results in Section IV. We close this paper with a conclusion in Section V.

II. Related Work

Large Language Models (LLMs) have shown remarkable adaptability across various domains, including legal texts, finance, and regulatory compliance, demonstrating their ability to detect inconsistencies, automate tasks, and improve efficiency [7], [11], [13], [14]. Recently, LLMs have emerged as a key force in advancing Named Entity Recognition (NER), as demonstrated in [15]–[20].

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

In the legal domain, LLMs like GPT-4 [1] and Llama-3 [2] have proven highly effective in handling the complexities of legal texts [21]–[23]. These models leverage their advanced language comprehension capabilities for tasks such as regulatory compliance verification [24]–[26], contract analysis [27], and legal violation detection [28]. A study on legal violation identification demonstrated that LLMs, in addition to BERT-based models, detected legal violations with a high degree of accuracy, outperforming traditional models like Conditional Random Fields (CRF) in legal contexts [28]–[30]. These models also handle long-range dependencies and context-rich information more effectively, leading to significant advancements in legal text processing and surpassing classical machine learning approaches [31], [32].

However, applying LLMs to legal NER remains challenging due to domain-specific terminology and intricate syntactic structures. Traditional NER models often underperform in legal contexts, as general language corpora fail to capture the nuances of legal texts. Domain-specific datasets, such as E-NER [33], LeNER-Br [29], and German-LER [34], have been developed to address these challenges by providing detailed annotations that significantly enhance NER performance across different legal systems and languages [28], [30], [35].

III. Methodology

This section sheds light on our approach to extracting named entities from legal documents and paragraphs.

A. Dataset Preparation

To evaluate the performance of each large language model (LLM), the primary goal is to assign IOB (Inside-Outside-Beginning) [36] tags to each sentence. We explore three strategies for querying LLMs: (1) directly retrieving IOB tags from the LLM, (2) obtaining tuples that specify the entity and its start and end positions, and (3) generating a JSON output where entities are represented as keys and their corresponding classifications as values.

The first two methods often lead to considerable inaccuracies. When LLMs return IOB tags directly, they handle tokenization internally, which frequently diverges from the dataset's ground truth tokenization, making the comparison between predicted and actual IOB tags unreliable. Similarly, in the case of returning start and end indices for entities, LLMs struggle to precisely identify the positions, and even slight discrepancies result in a complete misalignment of the IOB tags.

Given these limitations, we adopt the third approach. Rather than depending on the LLM to determine entity positions, we implement a code-based solution to accurately map entity positions within the sentence. This approach ensures a robust alignment with the ground truth, enabling a reliable evaluation of the IOB tagging performance.

B. Mapping of Entities to Sentence Tokens

To correctly map the model's returned entities (entity text) to sentence tokens, it is crucial to ensure proper alignment, as this directly impacts the accuracy of predicted IOB tags. Based on the dataset descriptions in relevant research, we replicate their tokenization methods for consistency and to allow for proper mapping. For instance, we use the NLTK [37] library for word tokenization in E-NER and leNER-br, SoMaJo [38] for German_LER, and Spacy[1] for InLegalNER. For other datasets, where the tokenization method is either unknown or completed manually by domain experts, we apply space-based tokenization. Since these datasets' sentences are pre-tokenized, we query the model with space-separated tokens and tokenize the returned entities similarly, using spaces. By adhering to these tokenization practices, we ensure accurate mapping of returned entity text to sentence tokens, resulting in reliable IOB tag predictions.

We also compare the consistency of re-tokenized sentences with the originals. In rare cases, discrepancies arise even with identical tokenization tools. To maintain accuracy in predictions and ground truth comparisons, we exclude such sentences. Additionally, when the same word belongs to different entities with varied classes, conflicts occur. In these cases, we retain the longer entity text (e.g., a company name containing a person's name) to resolve overlaps, ensuring that the more informative entity is preserved. Furthermore, large language models (LLMs) sometimes over-identify entities, producing entity classes absent from the dataset. In such cases, we disregard these extra entities to maintain dataset consistency.

C. Generation of JSON Output

During the generation phase, only three selected models (GPT-4o Mini [1], GPT-4o [1], and Mistral Large [39]) can enforce JSON outputs with their API calls. For models without this capability, we have to rely on the language model to produce a valid JSON output. If the model does not produce a valid JSON output, we evaluate this as if no entities are predicted.

D. Prompt Design

In our study, we design specific prompts for each dataset to ensure accurate entity extraction within distinct legal contexts. While the structure of the prompts remains consistent across datasets, we introduce dataset-specific instructions to optimize performance and address unique challenges inherent to each. For instance, the *Important Instructions* section consistently emphasizes three core requirements: maintaining a valid JSON format, strictly adhering to predefined entity classes, and handling ambiguities by excluding entities that do not match the provided categories.

However, we make tailored adjustments based on each dataset's specific requirements. In the *InLegalNER* dataset, titles or prefixes (e.g., "Mr.", "Sri.") are excluded from annotated entities, so we omit them from the extracted names. In contrast, in the *TurkishLegalNER* dataset, titles (e.g., "Tetkik Hakimi" or "Review Judge") are part of the PER class, so

[1] https://spacy.io/

we retain them to align with Turkish legal document norms. Similarly, in the *uk_ner_contracts* dataset, when extracting *Clause_Number*, we account for spaces between the number and period, ensuring any following periods are included in the entity. Datasets such as *LegalLensNER* focus on extracting violations and their legal context, requiring specific instructions to maintain consistency with legal terminology.

IV. Experiments

In this section, we describe our experimental protocol, examine the datasets and results, and discuss the strengths and limitations of our approach. All model training was performed on a shared GPU node featuring eight Nvidia V100 GPUs, an Intel Xeon 6148 CPU, and 1 TB of RAM.

A. Data

We use several legal NER datasets in this study. The first three, E-NER, InLegalNER, and LegalLensNER, are English-language datasets covering a variety of legal documents. The remaining datasets focus on other languages: leNER-br (Portuguese), German-LER (German), TurkishLegalNER (Turkish), and uk-ner-contracts (Ukrainian), each providing domain-specific annotations for their respective legal systems.

1) E-NER: The E-NER [33] dataset consists of 52 filings from the US SEC EDGAR database with manually annotated named entities. For this study, we select the version of the dataset that contains four entity classes: *Person*, *Organization*, *Location*, and *Miscellaneous*. Since the dataset does not come pre-split into training, validation, and test sets, we randomly select 20% of the data to be used as the test set.

2) InLegalNER: The InLegalNER [30] dataset, designed for legal NER in Indian legal texts, contains 46,545 annotated entities across 14 types, including *court names*, *petitioners*, *respondents*, and *statutes*. Since the dataset does not include IOB format annotations, we converted it by extracting the text and entity details (start/end positions, text, and entity class) for each sentence. Following the approach as outlined in their paper, we utilized spacy to map entity positions to corresponding words using the char_span method, labeling each word as the beginning (B-), inside (I-), or outside (O) of an entity. Partial manual verification showed that this method resulted in highly accurate mappings.

3) LegalLensNER: The LegalLensNER [28] dataset was initially generated by LLMs, but it has been carefully validated by expert annotators to ensure its accuracy and reliability. The dataset includes four main entity types: *Law*, *Violation*, *Violated By* (the entity committing the violation), and *Violated On* (the victim). For our study, we used the entire test set, which consists of 617 sentences.

However, one issue we identified with the dataset is the inconsistent labeling of the *Violated By* and *Violated On* entities. In some cases, unnecessary prepositions like “to” or “on” are included, while in others they are omitted. We believe that excluding these prepositions results in more accurate entity labeling. Instead of modifying the dataset, we ensured that the examples in our prompts avoided such inconsistencies, probably leading to a subjectively bad performance, but allowing for comparability to other approaches.

4) leNER-br: The leNER-Br [29] dataset was created for NER in Portuguese, specifically in Brazilian legal texts, containing manually annotated documents from various courts. It includes six entity types: *Pessoa* (persons), *Organizacao* (organizations), *Local* (locations), *Legislacao* (laws), *Jurisprudencia* (legal cases), and *Tempo* (time). This comprehensive dataset supports precise NER tasks in Portuguese legal documents. For our study, we used the entire test set, consisting of 1,389 sentences.

5) German-LER: The German Legal Entity Recognition (German-LER) [34] dataset is based on German legal texts and contains around 54,000 manually annotated entities. These entities are categorized into two types of classification: fine-grained and coarse-grained semantic classes. For our study, we selected the coarse-grained classification, which includes the following categories: *Person*, *Location*, *Organization*, *Legal norm*, *Regulation*, *Court decision*, and *Legal literature*. We used the entire test set, which consists of 6,673 sentences, for our experiments.

6) TurkishLegalNER: The TurkishLegalNER [32] dataset consists of annotated legal texts from the Turkish Court of Cassation, with a total of 2,198 sentences and 5,311 named entities. The dataset includes various entity types relevant to the legal domain, such as *PER* (Person), *LOC* (Location), *ORG* (Organization), DAT (Date), *LEG* (Legislation), *COU* (Court), *REF* (Reference), and *OFF* (Official Gazette). For our study, we used the test set, which contains 439 sentences.

However, due to privacy concerns, many sentences in the dataset contain “...” to obscure sensitive information. This particularly affects a significant portion of *PER* (Person) entities, as well as some *ORG* (Organization) and *LOC* (Location) entities. To mitigate this issue, we replace this anonymization technique with pseudo-anonymization, i.e., we replace these ellipses with appropriate terms, such as randomly selected Turkish names and locations, ensuring the dataset remains suitable for inference while adhering to privacy regulations.

7) uk-ner-contracts: The uk-ner-contracts [40] dataset classifies four key types of entities in Ukrainian legal contracts: Clause_number, Clause_title, Contract_type, and Definition_title. The dataset encompasses a wide range of legal documents across various domains, including employment, real estate, services, sales, and leases. All entities within the contracts have been manually labeled by legal experts, ensuring high-quality annotations. For our study, we use the test set, which consists of 494 sentences.

B. Evaluation Metrics

For evaluating the performance of LLMs, we use standard classification metrics based on a strict comparison of predicted IOB sequences with the ground truth labels. We enforce strict evaluation, requiring both the entity type and boundaries to exactly match the annotations, differing from implementation like [41].

TABLE I
FEW SHOT RESULTS

Model	**E-NER**	**InLegalNER**	**LegalLensNER**	**leNER-br**	**German-LER**	**TurkishLegalNER**	**uk-ner-contracts**
Dataset language	*English*	*English*	*English*	*Portuguese*	*German*	*Turkish*	*Ukrainian*
GPT-4o Mini	39.53	55.57	46.20	49.47	39.48	60.07	82.42
GPT-4o	48.24	**60.56**	49.65	**63.88**	57.00	**77.35**	**90.32**
Mistral [39]	37.86	58.39	44.82	51.09	43.33	68.78	88.48
Qwen2-72B [42]	**51.62**	51.76	45.19	59.01	52.07	70.41	75.00
Llama-3 70B [2]	44.25	59.40	38.28	61.87	**58.84**	66.12	20.14
Llama-3.1 8B [2]	25.44	42.54	35.28	43.31	28.71	23.37	14.48
Llama-3.1 70B [2]	40.98	51.46	44.87	48.95	44.78	52.11	15.77
Mixtral 8x7B [43]	32.28	33.97	26.00	35.70	26.36	20.08	9.54
Gemma-2 27B [44]	42.50	50.69	**50.08**	52.88	44.16	58.73	87.09
Phi-3 14B [3]	34.77	41.42	35.90	38.25	35.52	20.85	4.94
Phi-3 3.8B [3]	20.16	27.90	31.89	24.65	21.30	11.00	2.65

We computed precision, recall, and F_1 scores for each entity class to assess the model's performance. For reporting, we focused on the micro-averaged F_1 score, which aggregates the contributions of all classes, providing a balanced view of the LLM's overall accuracy in identifying and classifying entities, without being influenced by class imbalance. All results for the micro-averaged F_1 score were rounded to two decimal places.

C. Results

The evaluation of LLMs on several legal NER datasets revealed varied performance outcomes. As presented in Table I, F_1 scores for each model fluctuated across the seven datasets, reflecting differences in language and legal context.

GPT-4o [1] demonstrated the highest overall performance, consistently achieving top F_1 scores across multiple datasets. It excelled particularly in the *uk-ner-contracts* dataset, achieving an F_1 score of 90.32%, and showed strong results in *InLegalNER* and *German-LER*. Similarly, Qwen2-72B [42] and Mistral [39] achieved competitive outcomes, particularly in *InLegalNER* and *TurkishLegalNER*, underscoring their capability to manage complex legal texts across varied languages and systems.

In contrast, the Llama-3 and Llama-3.1 [2], usually a set of fairly strong models [12], displayed inconsistent performance across datasets, performing well on *leNER-br* with an F_1 score of 61.87% but struggling significantly on *uk-ner-contracts*, where it achieved only 20.14. This inconsistency suggests that while certain models excel in specific legal domains, their adaptability across datasets remains limited.

A deeper analysis of the *uk-ner-contracts* dataset highlighted specific challenges affecting model performance. The Clause Number entity class, comprising 80% of test cases, requires precise extraction for high accuracy. Clause numbers frequently appear in formats like ”1 .” or ”1.1.3 .”, where spacing and periods are integral to the entity. Despite clear prompt examples, many models struggled to reliably extract this entity, leading to performance discrepancies. However, models like GPT-4o and Mistral successfully adapted to these nuances, achieving superior results, which underscores the importance of precise formatting in legal NER tasks.

Furthermore, models like Mixtral 8x7B [43] and Phi-3 14B and 3.8B [3] struggled across most datasets, with notably low scores on *uk-ner-contracts* and *LegalLensNER*, highlighting the challenges that smaller models or less robust architectures face when tackling complex legal language and entity structures.

Overall, these results indicate that model architecture, dataset characteristics, and prompt design significantly influence performance in legal NER tasks. While proprietary models like GPT-4o set the standard, careful tuning and adaptation are crucial for consistent results across diverse legal datasets.

The substantial performance variation among LLMs on the *uk-ner-contracts* dataset is primarily due to the challenges associated with correctly identifying the *Clause_number* entity class. Among the four entity classes in this dataset (*Clause_number*, *Clause_title*, *Contract_type*, *Definition_title*), *Clause_number* is the most prevalent, constituting around 80% of test cases. Accurate identification of this entity is thus critical for achieving high overall performance.

In *uk-ner-contracts*, *Clause_number* entities commonly appear in formats such as “1 .” or “1.1.3 .,” where the final period and a preceding space are integral parts of the entity. The prompt examples explicitly illustrate this format and emphasize the importance of capturing these elements accurately. Despite this guidance, certain models fail to consistently recognize the space or final period, leading to substantial declines in accuracy. Conversely, other models like Gemma [44] or GPT-4o successfully interpret these examples, resulting in a strong performance on this dataset.

V. CONCLUSION

This study presents a comprehensive evaluation of a total of eleven large language models (LLMs) for Named Entity Recognition (NER) in the legal domain, covering seven diverse datasets in multiple languages and legal contexts. Our findings demonstrate that LLM performance varies significantly across legal NER tasks, influenced by model architecture and dataset-specific characteristics.

The proprietary GPT-4o [1] model consistently achieved the highest scores, particularly excelling in complex datasets like *uk-ner-contracts* due to its ability to adapt to intricate formatting requirements. Competitive results from models such as Qwen2-72B [42] and Mistral [39] underscore the potential of specialized LLMs to manage multilingual and nuanced legal texts. However, models like Llama-3.1 [2] and smaller architectures, including Mixtral 8x7B [43] and Phi-3 14B [3], showed limitations in handling the syntactic and structural complexities unique to legal language, resulting in inconsistent performance across datasets.

Future work could focus on developing adaptive prompt strategies and exploring domain-specific model architectures to further advance NER performance in specialized legal contexts. One could also explore various fine-tuning techniques to better capture the "legal language" of the datasets we investigated or improve the prompting strategy employed, as seen in e.g. [45] or [46].

Overall, this study highlights the substantial potential of LLMs for legal NER tasks, comparing the performances of many state-of-the-art LLMs and paving the way for other researchers to further improve on the baselines established in this study.

References

[1] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman *et al.*, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[2] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[3] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," 2024. [Online]. Available: https://arxiv.org/abs/2404.14219

[4] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," in *Proc. ICAIF*, 2023, pp. 349–356.

[5] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," 2023. [Online]. Available: https://arxiv.org/abs/2305.15005

[6] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "LLMs to the moon? Reddit market sentiment analysis with large language models," in *Proc. WWW*, 2023, p. 1014–1019.

[7] T. Deußer, D. Leonhard, L. Hillebrand, A. Berger, M. Khaled, S. Heiden, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa, "Uncovering inconsistencies and contradictions in financial reports using large language models," in *Proc. BigData*. IEEE, 2023, pp. 2814–2822.

[8] M. Pielka, S. Schmidt, and R. Sifa, "Generating prototypes for contradiction detection using large language models and linguistic rules," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 4684–4692.

[9] L. Xu, Z. Su, M. Yu, J. Xu, J. D. Choi, J. Zhou, and F. Liu, "Identifying factual inconsistencies in summaries: Grounding model inference via task taxonomy," 2024. [Online]. Available: https://arxiv.org/abs/2402.12821

[10] F. M. Plaza-del arco, D. Nozza, and D. Hovy, "Respectful or toxic? using zero-shot learning with language models to detect hate speech," in *The 7th Workshop on Online Abuse and Harms (WOAH)*, 2023, pp. 60–68.

[11] A. Berger, L. Hillebrand, D. Leonhard, T. Deußer, T. B. F. De Oliveira, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa, "Towards automated regulatory compliance verification in financial auditing with large language models," in *Proc. BigData*, 2023, pp. 4626–4635.

[12] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot arena: An open platform for evaluating llms by human preference," 2024.

[13] T. Deußer, M. Pielka, L. Pucknat, B. Jacob, T. Dilmaghani, M. Nourimand, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa, "Contradiction detection in financial reports," in *Proc. NLDL*, 2023.

[14] L. Hillebrand, A. Berger, T. Deußer, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, M. Pielka, D. Leonhard, C. Bauckhage, and R. Sifa, "Improving zero-shot text matching for financial auditing with large language models," in *Proc. DocEng*, 2023, pp. 1–4.

[15] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang, "Gpt-ner: Named entity recognition via large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.10428

[16] T. Deußer, L. Hillebrand, C. Bauckhage, and R. Sifa, "Informed named entity recognition decoding for generative language models," 2023.

[17] J. Li, H. Li, D. Sun, J. Wang, W. Zhang, Z. Wang, and G. Pan, "LLMs as bridges: Reformulating grounded multimodal named entity recognition," in *Findings of the ACL*, 2024, pp. 1302–1318.

[18] V. K. Keloth, Y. Hu, Q. Xie, X. Peng, Y. Wang, A. Zheng, M. Selek, K. Raja, C. H. Wei, Q. Jin *et al.*, "Advancing entity recognition in biomedicine via instruction tuning of large language models," *Bioinformatics*, vol. 40, no. 4, 2024.

[19] Y. Heng, C. Deng, Y. Li, Y. Yu, Y. Li, R. Zhang, and C. Zhang, "ProgGen: Generating named entity recognition datasets step-by-step with self-reflexive large language models," in *Findings of the ACL*, 2024.

[20] B. Subedi, S. Regmi, B. K. Bal, and P. Acharya, "Exploring the potential of large language models (LLMs) for low-resource languages: A study on named-entity recognition (NER) and part-of-speech (POS) tagging for Nepali language," in *Proc. LREC-COLING*, 2024.

[21] H. Jiang, X. Zhang, R. Mahari, D. Kessler, E. Ma, T. August, I. Li, A. Pentland, Y. Kim, D. Roy, and J. Kabbara, "Leveraging large language models for learning complex legal concepts through storytelling," in *Proc. ACL*, 2024.

[22] A. Deroy, K. Ghosh, and S. Ghosh, "Applicability of large language models and generative models for legal case judgement summarization," *Artificial Intelligence and Law*, pp. 1–44, 2024.

[23] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, and R. Perera, "Better call GPT, comparing large language models against lawyers," 2024. [Online]. Available: https://arxiv.org/abs/2401.16212

[24] A. Berger, L. Hillebrand, D. Leonhard, T. Deußer, T. B. F. De Oliveira, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, C. Bauckhage *et al.*, "Towards automated regulatory compliance verification in financial auditing with large language models," in *Proc. BigData*. IEEE, 2023, pp. 4626–4635.

[25] L. Hillebrand, M. Pielka, D. Leonhard, T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, M. Morad, C. Temath, T. Bell, R. Stenzel, and R. Sifa, "sustain.ai: a recommender system to analyze sustainability reports," in *Proc. ICAIL*, 2023, p. 412–416.

[26] J. Zhao and X. Wang, "Unleashing efficiency and insights: Exploring the potential applications and challenges of chatgpt in accounting," *J. of Corporate Accounting & Finance*, vol. 35, no. 1, pp. 269–276, 2024.

[27] K.-Y. Lam, V. C. Cheng, and Z. K. Yeong, "Applying large language models for enhancing contract drafting." in *Workshop LegalAIIA at ICAIL*, 2023, pp. 70–80.

[28] D. Bernsohn, G. Semo, Y. Vazana, G. Hayat, B. Hagag, J. Niklaus, R. Saha, and K. Truskovskyi, "LegalLens: Leveraging LLMs for legal violation identification in unstructured text," in *Proc EACL*, 2024.

[29] P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "LeNER-Br: a dataset for named entity recognition in Brazilian legal text," in *Proc. PROPOR*, 2018, pp. 313–323.

[30] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, and V. Raghavan, "Named entity recognition in Indian court judgments," in *Proc. NLLP Workshop*, 2022, pp. 184–193.

[31] V. Pais, M. Mitrofan, C. L. Gasan, V. Coneschi, and A. Ianov, "Named entity recognition in the Romanian legal domain," in *Proc. NLLP Workshop*, 2021, pp. 9–18.

[32] C. Çetindağ, B. Yazıcıoğlu, and A. Koç, "Named-entity recognition in turkish legal texts," *Natural Language Engineering*, vol. 29, no. 3, pp. 615–642, 2023.

[33] T. W. T. Au, V. Lampos, and I. Cox, "E-NER — an annotated named entity recognition corpus of legal text," in *Proc. NLLP Workshop*, 2022, pp. 246–255.

[34] E. Leitner, G. Rehm, and J. Moreno-Schneider, “A dataset of German legal documents for named entity recognition,” in *Proc. LREC*, May 2020, pp. 4478–4485.

[35] I. Angelidis, I. Chalkidis, and M. Koubarakis, “Named entity recognition, linking and generation for greek legislation,” in *Legal Knowledge and Information Systems*. IOS Press, 2018, pp. 1–10.

[36] L. Ramshaw and M. Marcus, “Text chunking using transformation-based learning,” in *Third Workshop on Very Large Corpora*, 1995. [Online]. Available: https://aclanthology.org/W95-0107

[37] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc., 2009.

[38] T. Proisl and P. Uhrig, “SoMaJo: State-of-the-art tokenization for German web and social media texts,” in *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, P. Cook, S. Evert, R. Schäfer, and E. Stemle, Eds., 2016, pp. 57–62.

[39] Mistral AI Team, “Mistral large,” February 2024, accessed: 2024-04-25. [Online]. Available: https://mistral.ai/news/mistral-large/

[40] LawInsider, “UK NER Contracts,” https://huggingface.co/datasets/lawinsider/uk_ner_contracts, accessed: 2024-10-24.

[41] T. Deußer, S. M. Ali, L. Hillebrand, D. Nurchalifah, B. Jacob, C. Bauckhage, and R. Sifa, “KPI-EDGAR: A novel dataset and accompanying metric for relation extraction from financial documents,” in *Proc. ICMLA*, 2022, pp. 1654–1659.

[42] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, “Qwen2 technical report,” 2024. [Online]. Available: https://arxiv.org/abs/2407.10671

[43] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand *et al.*, “Mixtral of experts,” 2024.

[44] G. Team, “Gemma,” 2024. [Online]. Available: https://www.kaggle.com/m/3301

[45] Y. Huang, K. Tang, and M. Chen, “Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies,” 2024. [Online]. Available: https://arxiv.org/abs/2402.09282

[46] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, “UniversalNER: Targeted distillation from large language models for open named entity recognition,” in *Proc. ICLR*, 2024.