

User-aware LLM-based Framework for Enhancing Contractual Clauses Compliance

Biagio Boi
dept. of Computer Science
University of Salerno
Fisciano, Salerno, Italy
bboi@unisa.it

Ilaria A. Caggiano
University of Naples
"Suor Orsola Benincasa"
Naples, Italy
ilaria.caggiano@unisob.na.it

Christian Esposito
dept. of Computer Science
University of Salerno
Fisciano, Salerno, Italy
esposito@unisa.it

Lucilla Gatt
University of Naples
"Suor Orsola Benincasa"
Naples, Italy
lucilla.gatt@unisob.na.it

Abstract—Contractual clarity is one of the most critical requirements in drafting a contract, yet its importance is often underestimated. This paper investigates the exploitability of Large Language Models (LLMs) to validate the comprehensiveness and correctness of contractual clauses, addressing the pressing need for precision in contract drafting, and the eventual possibility of getting improvement measures. Specifically, this study utilizes LLMs to automatically analyze and refine contractual clauses, focusing on improving clarity, coherence, and juridical compliance. By assessing the linguistic and structural features of the clauses, the LLM-based approach identifies potential sources of ambiguity and suggests modifications to ensure legal validity and reduce the risk of dispute. The framework's efficacy is demonstrated through a concrete case study composed of ten contractual clauses that illustrate its impact on enhancing contractual juridical compliance. This work represents a step forward in automating legal validation, offering a practical tool for creating legally robust contracts.

Index Terms—Legal document analysis, LLM, Contractual Clauses, NLP

I. INTRODUCTION

Current legislation mandates that contracts be written according to clear and correct clauses. Specifically, clauses drafted in an unclear and unintelligible manner may be classified as unfair or abusive and, therefore, invalid. These ambiguous and incomprehensible causes may cause a significant imbalance of the rights and obligations arising from the contract for the consumer, which can be debated in a dispute. This means that whenever a new contract is created, it is of pivotal importance to have linguistic correctness and clarity for the clauses in the contract. Moreover, it is important to have also the right references to law tools and articles. These checks are time-consuming and resource-intensive. Modern technologies might be leveraged to speed up the process and reduce the time needed to check the validity of contractual clauses. In various domain, as well as in the legal one, there is a recent debate on how good ICT technologies may perform in realizing some of the critical tasks. Therefore, this work aims to study the applicability of these technologies to the critical task of clauses validation and improvement.

Despite being widely used in assessing semantic correctness, Natural Language Processing (NLP) models cannot be directly leveraged in the case of contractual correctness. This is because correctness depends on current legislation,

which is frequently updated. Moreover, the legal language has intrinsic complexities, making translating the articles into a specific contractual clause difficult. In addition, a critical topic in modern legal scholarship is LegalAI: the role of automated tools in replacing paralegals, drafting contracts, and scrutinizing the terms of contracts [1]. While natural language models are increasingly capable of classifying contracts as clear or unclear, there remains a key question: how closely do these technologies align with human judgment, particularly in the legal domain? Tools leveraging Large Language Models (LLMs) offer significant advancements in contract interpretation, yet a gap persists between automated analysis and human interpretation [2].

In this work, we propose a tool to automatically extract clauses from a generic PDF contract and submit them to LLM-based tools for their validation and improvement. Within the context of LLMs, we can distinguish among two kinds: those that have been generally trained with no specific tasks, and we have considered ChatGPT [3], and those training for a specific kind of domain or task, and we have considered GiuriMatrix [4], a platform leveraging advanced AI and NLP technologies for specialized legal applications. ChatGPT represents a groundbreaking innovation in different fields, offering robust tools for automation, information management, and decision support. In contrast, GiuriMatrix is an AI-driven legal software designed to respond to regulatory queries in natural language. It serves as an assistant to legal professionals rather than a replacement, providing valuable support in daily legal practice. By utilizing ChatGPT and GiuriMatrix, we assess the effectiveness of AI in determining the correctness of contractual clauses and compare the results to human evaluations. In our analysis, we have integrated a human-in-the-loop concept [5] within the legal domain. This exploration highlights the potential benefits of integrating human expertise with automated tools in contract analysis.

The key contributions of this work are:

- Propose a novel framework that introduces human-in-the-loop analysis within the contractual correctness context, where the development of LLM-based solutions continues to accelerate, alongside a growing need for AI oversight;
- Evaluate the accuracy of LLMs in generating reliable responses regarding the semantic clarity of contracts

uploaded by users.

This paper is organized into five sections. Section two reviews related works in machine learning for legal applications. Section three presents the proposed framework. Section four discusses the results and gives insights into the accuracy of both statistical and LLM approaches to human clause interpretation. Finally, the paper concludes with insights into future developments in this field.

II. RELATED WORKS

In the legal domain, the primary focus of NLP research has been on information extraction and classification, leading to considerable efforts in developing reliable datasets for training models.

Chalkidis et al. [6] were among the first to introduce a large-scale text classification dataset for EU legislation. Later, Hendrycks et al. [7] proposed a dataset consisting of 500 contracts containing 41 different types of important clauses, with a total of 13,000 annotations. Additionally, *LexGLUE* [8] provides seven datasets for legal NLP tasks, including text classification for EUR-LEX, UNFAIR-ToS, and LEDGAR.

Notable applications of these datasets include the work by Xiao et al. [9], who introduced *Lawformer*, a pre-trained language model specifically designed for processing long Chinese legal documents. This model can manage documents with thousands of tokens, facilitating the development of legal judgment prediction datasets for both criminal and civil cases. *Lawformer* is based on *BERT* [10] and has been fine-tuned using a comprehensive dataset of 5,428,717 criminal cases and 17,387,874 civil cases, ensuring its suitability for legal data. Another notable application of *BERT* is seen in *BERT-PLI* [11], where the model was fine-tuned to capture semantic relationships at the paragraph level, enabling the inference of relevance between two legal cases through paragraph-level interactions. This approach aims to enhance legal case retrieval from generic text. Always in the Chinese legal perspective, Cui et al. [12] analyzed the problem of the knowledge in LLM, which is often incomplete or outdated, leading them to produce illusions that, although seemingly relevant, may be misleading or incorrect. They proposed a novel framework called *Chatlaw*, outperforming GPT4 by working on Retrieval-Augmented Generation (RAG) and by improving the attention mechanism behind the LLM.

We recognize the advancements made in the juridical domain; however, the analysis of contractual terms remains largely unexplored in the existing literature. Furthermore, despite the benefits introduced by these tools, practitioners continue to rely on general-purpose models such as ChatGPT. To address these gaps, we propose two Research Questions (RQs).

RQ1: How effective are LLMs, such as ChatGPT and GiuriMatrix, in evaluating the clarity of contractual clauses compared to traditional statistical tools like Dylan [13]?

The first RQ aims to assess the precision and accuracy of LLMs in comparison to statistical tools.

RQ2: What role does fine-tuning on juridical data (e.g., in the case of GiuriMatrix) play in enhancing LLMs' ability to accurately evaluate the clarity of contractual clauses?

The second RQ seeks to gain insights into the significance of fine-tuning, with a specific focus on the impact it has on the accuracy of general-purpose models compared to fine-tuned models.

III. METHODOLOGY

In this section, we present our solution and the methodologies employed at various stages in the development of this innovative framework. The primary goal of the framework is to automatically extract contractual clauses and then apply two distinct approaches to provide users with insights about the clarity of these clauses, helping them make informed decisions.

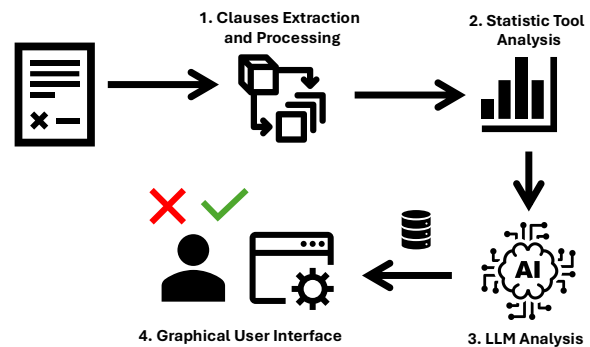


Fig. 1. Proposed Architecture

As shown in Fig. 1, the proposed architecture is structured into four phases:

- 1) **Clause Extraction and Processing** – This phase outlines the methods used to extract contractual clauses from legal documents and how they are processed for further analysis. This phase has been performed leveraging a PDF extractor such as the PyMuPDF library. Once the text has been extracted from the region of interest, it needs to be processed to identify and isolate specific clauses. This is achieved using regular expressions through Python's `re` library, which allows clauses to be detected based on predefined patterns.
- 2) **Statistic Tool Analysis** – Here, we describe the integration with Dylan, a statistic analysis tool, which is employed to evaluate the clarity of the clauses from a non-LLM perspective.
- 3) **LLM Analysis** – In this phase, we transfer the extracted clauses to LLMs to gain additional insights on the clarity and potential ambiguity of the clauses.

- 4) **Graphical User Interface** - After the evaluation process, we show the results to the user to make him responsible for the final decision applied over the clarity of the clause.

The identified clauses are then stored in an SQLite database for future use.

A. Statistic Tool Analysis

To obtain an automated evaluation of the clarity of the extracted clauses, a script was developed to interact with the Dylan API (sending the clauses in textual format and collecting readability scores for each clause), a tool for readability analysis. The results are then presented both to experts on the labeling page and used for further analysis.

Once extracted, each clause is sent to Dylan for readability analysis via a POST request. After Dylan processes a clause, the readability scores can be retrieved via a GET request using the document ID. This is handled by the `get_readability_scores()` function, which downloads the results in JSON format and organizes them into a dictionary. The main scores assigned by Dylan are:

- **Base:** Basic readability score of the text.
- **Lexical:** Score that measures the lexical complexity of the text.
- **Syntax:** Score related to the syntactic complexity of the text.
- **Global:** An overall evaluation that combines the base, lexical, and syntactic scores.

The function returns a dictionary containing the various readability scores, which are then printed for each clause. The script is executed by passing the path of a .docx or .txt file containing the contractual clauses. For each clause extracted from the file, it is uploaded to Dylan, followed by a request for readability scores and the printing of results. Each clause is processed individually, and the readability results are displayed to provide a detailed analysis of each contractual segment.

B. LLM Analysis

Unlike tools like Dylan, which focuses exclusively on basic linguistic aspects, Large Language Models (LLMs) can capture semantic and contextual nuances, which are particularly important in the legal domain. To improve the accuracy of the LLM model, we did not use it as a classical tool, but we opted for a prompt engineering process before to ask to the model our question. In particular, we entered the following prompt.

In the upcoming interactions, I will send you some clauses, and you will need to evaluate their clarity by giving me a score from 1 to 5, where 1 is unclear and 5 is very clear.

The goal is to obtain a quantitative as well as a qualitative evaluation (based on how the two models justify their assessments) of the clarity of the clauses, similar to the ratings provided by human participants and the Dylan tool. This

secondary analysis aims to verify whether advanced language models, with their ability to understand semantic and legal context, can offer an assessment closer to human judgment compared to purely statistical metrics.

In the legal field, a clause may be considered clear not because it is brief or contains simple terms, but because it leaves no room for interpretive ambiguity. Thus, semantic analysis plays a central role: clarity does not depend solely on word complexity but on its appropriate use within the legal context. An LLM, such as ChatGPT or GiuriMatrix, has the potential to provide evaluations that go beyond statistical metrics, constructing semantic relationships that cannot be captured by a purely linguistic algorithm.

C. Graphical User Interface

The readability scores are displayed to the users so that the user can compare it own evaluation with respect to the evaluation performed with the model. After obtaining clarity scores from both the statistical tools and the LLM, we developed a Graphical User Interface (GUI) to facilitate human-in-the-loop decision-making processes. To achieve this, we implemented a web platform designed to present all the scores generated by the statistical and LLM analyses. For each clause, the interface displays the suggestions provided by the statistical tool alongside those from the LLM, as illustrated in Fig. 2.

Clause Analysis			
Clause	<input type="radio"/> Readable	Dylan Response	LLM Response
	<input checked="" type="radio"/> Not Readable		
The Client declares themselves a debtor towards the Financing Company for the amount eventually requested as Pre-financing, which they agree to repay in a single installment, withheld from	<input type="radio"/>	Base: 54.9 Lexical: 84.8 Syntactic: 99.3 Global: 99.6	58.00

Fig. 2. GUI shown to a typical user

The clarity score is measured on a scale from 0 to 100, where 0 indicates highly readable text, and 100 denotes high levels of ambiguity and semantic complexity.

The final judgment is always entrusted to the expert, who retains full responsibility for deciding whether a clause is readable or not. The GUI presents the results from the analysis tools, allowing the expert to classify each clause accordingly in a dedicated column. The expert's classifications are stored in a database, enabling future reuse for customized fine-tuning of the LLM and improving the assessment of these tools' accuracy.

IV. RESULTS

We will now evaluate the proposed methodology through the developed platform. On the basis of extracted contractual clauses thought the usage of the proposed platform, we created

a set of ten clauses, on which we created a questionnaire to validate the response of both Dylan and two different LLM models.

The participants have been splitted into three clusters, as depicted in Tab. I. These three clusters are splitted into:

- 1) **Students, in both humanities or sciences disciplines** - A group of students who do not have a specific juridic formation, but come from different faculties.
- 2) **Legal experts**: Legal professionals, such as lawyers, accountants, and notaries, with extensive training and professional practice in the field.
- 3) **Non-legal experts**: Participants without specific training or experience in the legal field.

TABLE I
SUMMARY OF AGE GROUPS, FIELDS OF STUDY, AND ROLES OF PERSONS WHO ANSWERED THE QUESTIONNAIRE.

Category	Group	Percentage (%)
Age Group	18-30	66.70
	31-50	23.30
	51-70	10.00
Field of Study	Humanities	33.30
	Sciences	66.70
Role	Student	50.00
	Legal Expert	20.00
	Non-legal Expert	30.00
Expertise	Lawyer	33.30
	None of them	66.70

A. Questionnaire

To validate the response from Dylan and from the involved LLMs, we created a questionnaire composed of ten questions, corresponding to the ten clauses extracted from the previous point. We used Google Forms to automatize the collection of the responses. The current version of the questionnaire can be reached at the following link: <https://forms.gle/v6i9xZmvoKsKuTcu8>, shared among different social media.

The questionnaire presents the extracted clauses and asks participants to evaluate each one as *Unclear* or *Very clear* on a scale from 1 to 5, respectively. Demographic questions are included to distinguish between students, experts, and non-experts.

In Tab. II, we represented the average responses to the questionnaire. The average score exceeds 2.5, suggesting that the clauses are generally considered clear by the participants. However, legal experts assigned significantly lower scores to clauses 1, 2, and 9, while non-legal experts rated these clauses as having higher clarity. This discrepancy may indicate that non-experts struggle to detect ambiguity, leading them to perceive certain clauses as clearer than they truly are.

Legal experts—both practicing professionals (such as lawyers) and those with legal training who do not practice directly—tended to rate the clauses as clearer compared to non-experts. This reflects their ability to understand the technical language and structure of contracts.

B. Statistic Tool Performance

To compare the results between the questionnaire and the statistical tool, we needed to compute a consistent score across both methods. Since Dylan evaluations are inversely proportional to the participants responses and use a different scoring mechanism, we converted the questionnaire values using the following formula:

$$P(x) = (5 - x)20 \quad (1)$$

where x is the average response score given by participants for a particular clause. This formula inverts the scale by subtracting the response score from 5, and then multiplying by 20 to reflect the proportionate scaling. As a result, a lower original score corresponds to a higher percentage (up to 100%), while a higher score corresponds to a lower percentage (down to 0%), aligning with Dylan's scoring approach.

It is important to note that this comparison involves two distinct scoring systems: on one hand, the subjective evaluations, which may be influenced by participants' educational backgrounds and the specific meaning of the sentences; on the other hand, the statistical tool, which strictly measures lexical complexity without accounting for contextual understanding.

This approach allows us to analyze the differences between human perception and the parameters used by an automated system to measure readability. The data reveal a significant discrepancy between the two approaches. While Dylan's evaluations for the overall score of each clause exceed 90%, indicating a very high level of difficulty, human evaluations do not surpass 44%. This imbalance arises from the nature of Dylan, which is designed to assess linguistic simplicity.

The legal context differs markedly from purely linguistic analysis. Tools mentioned in the literature, such as Dylan, were originally developed to measure linguistic simplicity in a more general setting, such as educational contexts (e.g., evaluating texts intended for elementary or high school students). These tools focus on linguistic factors such as sentence length, lexical complexity, and grammatical structure—elements that are naturally more complex in contractual texts than in other forms of writing. In practice, Dylan fails to capture the semantic understanding that a legal expert or a student with strong interpretive skills might possess, limiting its evaluation to the linguistic form of the text.

C. LLM Performance

In the second phase of the analysis, two advanced language models will be used to evaluate the same set of clauses, namely ChatGPT and GiuriMatrix.

ChatGPT may rate a clause as very difficult, while GiuriMatrix, due to its targeted training, might consider it clearer. This comparison will help determine whether a generic LLM like ChatGPT is sufficient for such tasks, or if specialized models like GiuriMatrix are preferable in complex contexts such as the legal field. The evaluation process in this phase will be guided by prompt engineering, i.e., formulating specific requests to the language models to achieve a consistent assessment of the contractual clauses. The clauses extracted during the first

TABLE II
AVERAGE RESPONSE FOR EACH USER TYPE AND CONTRACTUAL CLAUSE GIVEN TO THE PROPOSED QUESTIONNAIRE.

User Type	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
Legal experts: Lawyer	1.50	1.75	4.50	4.75	3.00	4.75	4.00	4.50	2.75	4.25
Legal experts: Generic	4.00	3.00	3.50	3.00	3.00	4.50	4.50	4.00	4.00	3.50
Non-legal experts	2.78	3.56	4.00	3.89	3.44	3.67	3.89	3.67	3.89	4.22
Students: Humanities	4.00	3.40	3.60	3.80	4.20	4.80	4.60	4.40	4.00	4.00
Students: Science	2.50	3.10	3.00	3.40	2.50	3.70	3.30	2.90	3.20	3.20
Average	2.80	3.10	3.63	3.77	3.17	4.07	3.87	3.67	3.60	3.80

phase will be submitted to both ChatGPT and GiuriMatrix with identical prompts. The first significant difference between the two models emerges from the way they respond to the proposed prompt, as discussed in the previous section.

- **ChatGPT:** After receiving the initial prompt, ChatGPT understands the context and the request, enabling it to process each subsequent clause without requiring further instructions. This means that once the context is set with the first prompt, ChatGPT responds directly with a clarity evaluation without needing to repeat the command for each clause.
- **GiuriMatrix:** In contrast, GiuriMatrix requires additional instructions for every interaction. Each time I submit a clause, I must explicitly restate what I want it to do (i.e., evaluate clarity and return a score from 1 to 5). Although GiuriMatrix can analyze the clauses, it does not seem to maintain a stable context after the first prompt, implying that each new clause must be accompanied by a repetition of the command.

This difference may seem minor at first glance, but it significantly impacts the fluidity of interaction. Below, we present an example of an answer given from the two models, showing the clarity evaluations and comparing the responses provided by the two language models (ChatGPT and GiuriMatrix) with those of the statistical tool Dylan. This comparison allows us to observe the differences between automated assessments, considering both Dylan’s statistical linguistic perspective and the semantic and contextual understanding of the LLMs.

Response from ChatGPT

*The clause is quite complex and includes several financial elements and specific terms that may not be immediately clear to those who are not familiar with this type of contractual language. However, the sentence structure and the connections between the various parts are logical, albeit dense with information. More readings may be necessary for a complete understanding, especially for those who are not experts in the field. **Clarity score: 3/5.***

To improve clarity, I would suggest breaking the clause into several shorter sentences, better outlining the conditions and consequences, and further clarifying the implications of the lack of funding.

Response from GiuriMatrix

*The provided clause is fairly clear and detailed. The conditions of the pre-financing, including the terms of repayment and interest, are specified clearly. **I would rate this clause with a score of 4.***

From the analysis of the responses provided by the two language models, a significant difference emerged in how ChatGPT and GiuriMatrix approach the evaluation of the clarity of contractual clauses. ChatGPT stands out for its greater comprehensiveness in responses. In addition to evaluating the clarity of the clause on a scale from 1 to 5, it provides an explanation and suggestions for improving the clause’s formulation to make it clearer and more understandable. These suggestions include advice on linguistic, structural, and lexical modifications, demonstrating a broader and deeper reflective capability. Its generic nature, while not specifically trained on legal texts, allows it to analyze the text with a certain flexibility, attempting to resolve ambiguities that may affect comprehension.

Conversely, GiuriMatrix tends to be more concise in its responses, offering shorter and less articulated evaluations without providing suggestions for improving the clauses themselves. This may reflect its specialization in the legal field, where precision and efficiency are prioritized over broader analysis or rewriting recommendations.

D. Discussion

Dylan, as a linguistic-statistical tool, has several limitations when applied to legal texts. While human users evaluate clauses based on their overall understanding, considering context and semantics, Dylan cannot do this. Its analysis is limited to the length of sentences and words, the frequency of certain words compared to a common lexicon, and to the grammatical complexity of sentences.

As a result, Dylan consistently assigns very high difficulty scores to contractual texts, as it cannot recognize the conceptual clarity or legal intent that is often understood by human experts. Its assessment tends to be particularly penalized by the length of sentences and the presence of uncommon technical words, ignoring the fact that these clauses, while formally complex, may be perfectly comprehensible to a lawyer or an expert, if not even to a student, as evidenced by the results.

Therefore, while Dylan provides useful information for the *technical readability* of a clause, it is unable to accurately

assess its overall clarity from a semantic perspective. This is likely due to the fact that the *training corpus* is composed of two corpora belonging to the same textual genre (journalistic prose): La Repubblica (Rep) and Due Parole (2Par); the latter consists of articles written for an audience of adults with cognitive deficits or characterized by a low level of literacy. It would therefore be advisable to create a basic vocabulary in the legal field that would allow for refining the tool.

Human evaluations, especially those from legal experts, remain essential for a comprehensive and thorough analysis of the clarity of contractual clauses. To address this issue, we take into account the use of LLMs, which could offer a more advanced assessment of textual clarity, going beyond statistical indices.

Supported by the results, we can give a response to RQ1 by assessing that the evaluations provided by ChatGPT and GiuriMatrix are much closer to those obtained from human users, whereas the statistical tool Dylan tends to diverge significantly from both human assessments and those of the two language models. This discrepancy arises because artificial intelligence models like ChatGPT and GiuriMatrix can represent a range of evaluations more aligned with human perception, as they can manage and understand text in a more contextual and semantic manner. Indeed, LLMs tend to provide evaluations that align with the average of the human responses collected through the questionnaire, as they represent a more complex and flexible interpretation of language, akin to what a sample of individuals would provide. In contrast, Dylan, as a statistical tool focused on purely linguistic parameters (such as sentence length and lexical complexity), tends to rate clauses as very difficult in almost all cases, as it is not designed to capture the nuances and specificities of the legal context. Thus, LLMs represent more suitable and promising tools for evaluating the clarity of clauses. This leads to the conclusion that statistical tools are not suitable for readability analysis in legal contexts, whereas LLMs undoubtedly serve as a strong starting point for developing increasingly effective tools in this field.

Difference in answering to the same question from ChatGPT and GiuriMatrix let us give a substantial answer to the RQ2 concerning the impact of fine-tuning on legal data on the models' ability to accurately judge the clarity of legal texts. GiuriMatrix tends to rate many clauses as clearer compared to ChatGPT, with an average difference of about one point on the evaluation scale. This may be explained by the fact that GiuriMatrix has undergone a fine-tuning process specific to the legal domain, enabling it to interpret contractual clauses with greater familiarity and recognize structures that might appear more complex to ChatGPT due to its lack of specialization in this area.

V. CONCLUSION

This study presented an approach to assessing contractual clarity through a web-based platform that leverages recent advancements in LLMs and NLP. While LLMs are often misunderstood within the legal community, this research demonstrates their superior efficacy compared to traditional statistical

tools, which frequently fail to comprehensively analyze legal texts. The proposed framework paves the way for a collaborative environment between legal experts and ML models. A limit of this study is related to the corpus of ten clauses being used. In future work, we aim to enlarge this dataset. We were affected by the legal restrictions of having access to contractual clauses debated in disputes and the fact that web-hosted contracts are very similar for banking and insurance products. To overcome such a limitation, we aim to generate synthetic data to have a larger dataset to train and test our solution. We have used ChatGPT only, but we plan to enlarge our investigation to other LLMs, such as Gemini or LLama3.

ACKNOWLEDGMENTS

The paper is the result of the work of a selected team of researchers of the Research Centre of European Private Law (ReCEPL) coordinated by the ReCEPL Director Prof. Lucilla Gatt.

REFERENCES

- [1] D. Bhora and K. Shrahan, "Demystifying the role of artificial intelligence in legal practice," *Nirma ULJ*, vol. 8, p. 1, 2018.
- [2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does nlp benefit legal system: A summary of legal artificial intelligence," *arXiv preprint arXiv:2004.12158*, 2020.
- [3] J. Deng and Y. Lin, "The benefits and challenges of chatgpt: An overview," *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 2, pp. 81–83, 2022.
- [4] M. Filippelli *et al.*, "Giurimatrix," <http://www.giurimatrix.it/>, 2021.
- [5] M. Almada, "Human intervention in automated decision-making: Toward the construction of contestable systems," in *Proceedings of the Seventeenth International Conference on artificial intelligence and law*, 2019, pp. 2–11.
- [6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on eu legislation," *arXiv preprint arXiv:1906.02192*, 2019.
- [7] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "Cuad: An expert-annotated nlp dataset for legal contract review," *arXiv preprint arXiv:2103.06268*, 2021.
- [8] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "Lexglue: A benchmark dataset for legal language understanding in english," *arXiv preprint arXiv:2110.00976*, 2021.
- [9] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021.
- [10] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.
- [11] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma, "Bert-ppl: Modeling paragraph-level interactions for legal case retrieval," in *IJCAI*, 2020, pp. 3501–3507.
- [12] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan, "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," *arXiv preprint arXiv:2306.16092*, 2024.
- [13] F. Dell'Orletta, S. Montemagni, and G. Venturi, "Read-it: Assessing readability of italian texts with a view to text simplification," in *Proceedings of the second workshop on speech and language processing for assistive technologies*, 2011, pp. 73–83.