On Exploiting LLMs and Statistical Methods for Testing Clarity in Legal Contracts

Biagio Boi
Christian Esposito
University of Salerno
Fisciano, Italy
{bboi,esposito}@unisa.it

Ilaria A. Caggiano
Lucilla Gatt
University of Naples "Suor Orsola Benincasa"
Napoli, Italy
{ilaria.caggiano,lucilla.gatt}@unisob.it

Abstract

Current legislation requires contracts to be written clearly and concisely. However, many contracts remain ambiguous and challenging for readers to understand. Advancements in natural language analysis using statistical and Large Language Models (LLMs) are improving the process of clarity verification by reducing the time needed for the overall process. In this paper, we investigate the potential of LLMs, such as ChatGPT and Giuri-Matrix, against existing statistical tools for natural language clarity checks. Results suggest the adaptability of traditional LLMs in verifying contractual clarity and providing suggestions for improvement of submitted contracts.

CCS Concepts

• **Applied computing** → **Law**; *Document analysis*; • **Computing methodologies** → *Neural networks*.

Keywords

Large Language Model (LLM), Natural Language Processing (NLP), Contractual Clarity

ACM Reference Format:
Biagio Boi, Christian Esposito, Ilaria A. Caggiano, and Lucilla Gatt. 2025. On Exploiting LLMs and Statistical Methods for Testing Clarity in Legal Contracts. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), March 31-April 4, 2025, Catania, Italy*. ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3672608.3707955

1 Introduction

Although current technological progress in Natural Language Processing (NLP) and Large Language Models (LLM) has made notable strides in assessing whether texts are clear or not, a critical question arises from the legal context [1, 3]: *How accurately do these technologies replicate human judgment, especially within the legal field?* Statistical tools have long provided a reliable framework for identifying patterns and trends in legal texts, particularly in assessing clause precision. Recently, LLMs have emerged as a powerful alternative, offering sophisticated capabilities in understanding and generating human language. This study compares the reliability of these two methods to determine which is more effective in assessing contractual clarity and reducing disputes. Specifically, we consider two LLMs: ChatGPT, a general-purpose assistant, and GiuriMatrix, a platform that leverages advanced AI and NLP technologies for specialized legal applications. Fine-tuning is a relevant process for improving the quality of LLMs in a particular context [2]; however, this is not free, requiring additional computation and, therefore, costs. This motivates our research, since all the contexts are the same, and it might be possible that the juridical one is one of these. Perhaps the role of LLMs in a legal context is unclear, where complex terms are typically used. Furthermore, despite the benefits introduced by literature tools, practitioners continue to rely on general-purpose models such as ChatGPT, which motivate the research to understand its role in a juridical context better.

2 Methodology

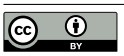
To carry out our experiments, we extracted a set of ten clauses from a contract to a salary-backed loan provided by an insurance company, in which we created a questionnaire to validate the responses of both the statistical tool and two different LLM models. The participants have been split into three clusters:

(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.

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 available at https://forms.gle/v6i9xZmvoKsKuTcu8. In Tab. 1, we represented the average responses to the questionnaire. We needed to compute a consistent score across both methods to compare the results between the questionnaire and the statistical tool. Since Dylan's evaluations are inversely proportional to the participant's responses and use a different scoring mechanism, we converted the questionnaire values using the formula

$$P(x) = (5 - x) \times 20$$

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 multiplying by 20 to reflect the proportionate scaling. As a result, a lower original score corresponds to a higher percentage (up to 100%), reflecting the huge complexity, while a higher score corresponds to a lower rate (down to 0%), reflecting the clarity of clauses; both aligning with Dylan's scoring



SAC '25, March 31-April 4, 2025, Catania, Italy

ACM ISBN 979-8-4007-0629-5/25/03
https://doi.org/10.1145/3672608.3707955

Table 1: 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 Clarity	**2.80**	**3.10**	**3.63**	**3.77**	**3.17**	**4.07**	**3.87**	**3.67**	**3.60**	**3.80**
Average Complex. (%)	**44.00**	**38.00**	**27.40**	**24.60**	**36.60**	**18.60**	**22.60**	**26.60**	**28.00**	**24.00**
Statistical Tool Complex. (%)	**54.90**	**30.50**	**74.50**	**59.00**	**53.60**	**64.40**	**68.00**	**70.40**	**38.70**	**85.30**
LLM - ChatGPT - Complexity (%)	**40.00**	**60.00**	**40.00**	**60.00**	**55.00**	**70.00**	**50.00**	**65.00**	**60.00**	**55.00**
LLM - GiuriMatrix - Complexity (%)	**20.00**	**20.00**	**20.00**	**30.00**	**40.00**	**35.00**	**45.00**	**25.00**	**40.00**	**30.00**

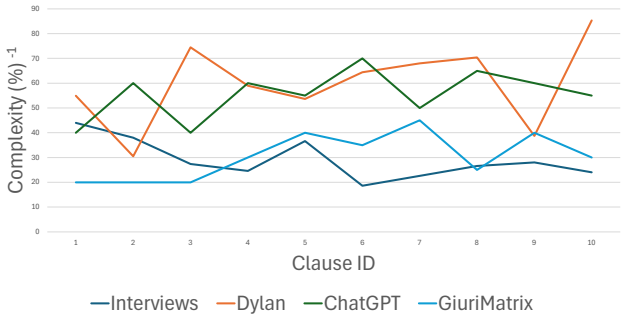


Figure 1: Complexity trend over the analyzed clauses

approach. The data reveal a significant discrepancy between the two methods. While Dylan's evaluations for the overall score of each clause exceed 50% for different clauses, human evaluations do not surpass 44% in any of them. This imbalance arises from the nature of Dylan, which is designed to assess linguistic simplicity.

The two LLMs were used in the second phase of the analysis. The first significant difference between the two models emerges from how they respond to the proposed prompt. As in Tab. 1 and Fig. 1, ChatGPT is closer to Dylan for most of the clauses, while GiuriMatrix follows the trend of the interviewed people. In addition to evaluating the clause complexity from 0 to 100, ChatGPT also provides an explanation and suggestions for improving the clause to make it more straightforward and understandable. Conversely, GiuriMatrix tends to be more concise in its responses, offering shorter and less articulated evaluations without providing suggestions for improving the clauses. This may reflect its specialization in the legal field, where precision and efficiency are prioritized over broader analysis or rewriting recommendations.

3 Discussion & Conclusion

Dylan consistently assigns high scores to contractual texts because it fails to recognize the conceptual clarity and legal intent that human experts readily understand. Its assessments are disproportionately influenced by sentence length and the use of uncommon technical terms, overlooking the fact that these clauses, despite their formal complexity, can be perfectly comprehensible to lawyers, legal experts, and even students, as demonstrated by the results. To overcome this limitation, we explore the use of LLMs, which can provide a more nuanced evaluation of textual clarity that extends beyond traditional statistical indices. The results support our research question: ChatGPT delivers a more qualitative assessment, and despite the reasoning behind its evaluation, the final scores align with those of the statistical tool. In contrast, GiuriMatrix produces a more precise score, reflecting its specialized ability to interpret legal texts. Scores of this last tool are much more closer to the human perception

Thus, LLMs represent more suitable and promising tools for evaluating the clarity of clauses, if fine-tuning is applied in a consistent manner. This leads to the conclusion that statistical tools are unsuitable for legal readability analysis. In contrast, LLMs undoubtedly serve as a strong starting point for developing increasingly effective tools in this field; where fine-tuning is needed to make the LLM able to better understand the specific context. Looking ahead, we aim to develop a specialized LLM that is explicitly fine-tuned for contractual clauses. This model will offer deeper insights and incorporate principles of explainable AI, which is increasingly demanded in contemporary applications.

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] Darshan Bhora and Kuldeep Shravan. 2018. Demystifying the role of artificial intelligence in legal practice. *Nirma ULJ* 8 (2018), 1.
[2] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. arXiv:2402.17193 [cs.CL] https://arxiv.org/abs/2402.17193
[3] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. arXiv:2004.12158 [cs.CL] https://arxiv.org/abs/2004.12158