Towards LLM-based Semantic Analysis of Historical Legal Documents

Tania Litaina[1], Andreas Soularidis[1,*], Georgios Bouchouras[1,*], Konstantinos Kotis[1] and Evangelia Kavakli[1]

[1]Dept. of Cultural Technology and Communication University of Aegean, University Hill, Mytilene, Greece

Abstract

The preservation of legal documents such as notarial ones is of vital importance as they are evidence of legal transactions between the involved entities through the years, serving as historical legal knowledge bases. The emergence of Large Language Models (LLMs) and their ability to analyze big data and generate content (much faster and relatively better than humans do alone) has created new perspectives in many fields, including law. Motivated by the significant potential of LLMs, we investigate the capabilities and limitations of using them in semantically analyzing legal documents through experimentation with two most prevalent LLMs i.e., ChatGPT-3.5 and Gemini/Bard. The goal is to emphasize automated and faster semantic analysis of documents, placing questions (prompts) concerning the type and subject of contracts, the recognition of the involved named entities and their relationship(s) e.g., landlord-tenant or family relationships. The experiments conducted with digitized contract documents that have been converted from handwritten Greek originals into plain text (LLM input) using Transkribus, an AI-powered platform for text recognition and transcription. The LLM responses were evaluated against the results obtained from a human expert, performing better in terms of precision but not in recall.

Keywords

LLM, legal documents, semantic analysis, named entity recognition

1. Introduction

Legal documents such as notarial acts serve as records of legal transactions over time, documenting various human activities and relationships, constituting historical legal knowledge bases. The advent of digitization has facilitated the long-term preservation and safeguarding of information, enabling wider dissemination and access to the digitized documents. However, automated content analysis of the digitized documents remains a key challenge especially when the extraction of information is performed by the transcription of old manuscripts in languages other than English, Greek in our case, where the visual recognition of characters in the text can be more complicated. In this context, Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Learning (ML) algorithms have been proven very important [1].

Large Language Models (LLMs) are models trained on a massive amount of data demonstrating remarkable ability to analyze data and generate human-like text bringing significant changes in

SemDH2024: First International Workshop of Semantic Digital Humanities, co-located with ESWC2024, May 26 27, 2024, Hersonissos, Greece

ctm22013@ct.aegean.gr (T. Litaina); soularidis@aegean.gr (A. Soularidis); cti23010@ct.aegean.gr (G. Bouchouras); kotis@aegean.gr (K. Kotis); kavakli@aegean.gr (E. Kavakli)

0009-0006-9769-5583 (A. Soularidis); 0000-0003-0566-3615 (G. Bouchouras); 0000-0001-7838-9691 (K. Kotis); 0000-0003-2743-5146 (E. Kavakli)

everyday life due to their increasing contribution across diverse domains including law. The contribution of such tools to the analysis of legal documents paves the way for automation and improvements of legal processes, facilitating time-consuming tasks [2]. Utilizing LLMs within the legal domain necessitates the examination of vast quantities of relevant data, such as notarial documents and legal texts. This poses a significant challenge, particularly when tailored for diverse languages and national legal frameworks.

In this context this paper reports initial results regarding the effectiveness of LLMs for the semantic analysis of handwritten contractual deeds of the late 19th century from the Greek State Archives digital collection. The selection of the documents was based on their historical significance, as well as several methodological challenges arising from the utilization of calligraphy and the orthographic peculiarities of the scribe which aggravate the recognition of many words. Initially, the scanned documents have been transcribed using Transkribus [3], a machine learning platform that has been trained to recognize handwritten text. The quality of the transcription was then assessed by domain experts (archivists). While this has significantly enhanced the legibility of the text, further difficulties arose in comprehending and analyzing it, mainly because the documents are written in "katharevousa", a conservative and literary variety of modern Greek, which until 1976 was used in government and judiciary documents. After the transcription process, the generated plain texts were used as input for the selected LLMs. To this end, we have experimented with two widely used and freely available LLMs: ChatGPT-3.5[1] and Gemini/Bard[2]. Our aim was to investigate the capabilities and weaknesses of LLMs when working with such types of documents, ultimately leading to the definition of a complete methodology for the automated processing and analysis of old handwritten Greek legal documents based on AI.

The structure of this short paper is as follows: Section 2 presents the related work regarding LLMs and their use in the semantic analysis of legal documents. Section 3 describes the research methodology, while section 4 reports the conducted experiments. Section 5 discusses the results, and finally, section 6 concludes the paper with pointers to future work.

2. Related Work

LLMs have already shown promising results in legal document analysis, contract review, and legal research. A number of studies emphasize the legal problems that arise from the integration of LLMs in the legal field such as intellectual property, data privacy and bias [4]. In this paper we focus on evaluating the efficacy of LLMs in comprehending legal texts.

In more detail, Blair-Stanek et al. [5] investigate to what extend LLMs are capable of performing statutory reasoning, one of the most basic tasks required by lawyers. Towards this direction they use GPT-3 model and the U.S tax related StAtutory Reasoning Assessment (SARA) dataset. The study investigates the impact of various prompting approaches to further improve the reasoning capabilities. The experimental results demonstrate the reasoning ability of GPT-3 with statutes, highlighting also errors due to limited prior knowledge.

Nay et al. [6] investigate the capabilities of LLMs in understanding the tax law, an area

[1]https://chat.openai.com/
[2]https://gemini.google.com/app

that requires reasoning and math skills. They validate the understanding capabilities of LLMs (ChatGPT-3, ChatGPT-3.5, ChatGPT-4) using multiple-choice legal problems based on tax law, while they assess the performance of LLMs implementing various experimental setups that include few-shot prompting, presenting examples of question-answer pairs, and integrating them with legal text from U.S federation using various retrieval techniques. The experimental results demonstrate the ability of LLMs to perform at high levels of accuracy, especially using enhanced prompt approaches, but still far from the tax lawyer levels.

Fei et al. [7] propose a comprehensive evaluation benchmark, namely LawBench. They assess 51 LLMs in legal capabilities such as knowledge a) memorization, b) understanding, and c) application. The experimental results demonstrate the superiority of ChatGPT-4 in the law domain, generating better results even from fine-tuning LLMs on legal specific language. However, the results also highlight the weakness of current LLMs to provide meaningful aid in the law domain.

The limitations highlighted by the above studies concerning the ability of LLMs to comprehend and analyze legal texts, underscore the need to further explore capabilities along a wider range of legal documents, which is not limited solely to federal laws and English-written documents

3. Research methodology

The research presented in this paper follows a four-phased approach, to systematically uncover and assess the capabilities and limitations of LLMs, using ChatGPT-3.5 and Gemini/Bard, in the context of semantic analysis, and comprehension of historical Greek contracts (summarized in Figure 1). Regarding the notarial documents used in the experiments, initial samples were obtained from the internet in both English and Greek languages, whilst the main experimentation corpus was comprised of a representative set of 17 handwritten contracts of the 19th century which were initially transcribed using the Transkribus platform and selected due to their quality of the transcription. To evaluate the responses obtained by the LLMs, these were compared to the ones of a human expert measuring precision and recall.

At the first stage, experiments were conducted using notarial documents written in English. The aim was to investigate whether and to what extent ChatGPT-3.5, as well as LLMs in general, are able to understand a notarial document, identifying entities, relationships between them, or even predicting an upcoming/future legal act.

Similar analysis was pursued in the second phase, but this time both ChatGPT-3.5 and Gemini/Bard were tested. More specifically, a ”declaration of acceptance of inheritance”, written in modern Greek was used, which concerns the transfer of property from a father to his son and the transfer of son’s property to his own children. Afterwards, the same document was manually translated into English to investigate whether the LLMs generate more accurate results compared to the Greek version. This time the prompts were more constrained, and emphasis was given on identifying relationships between the contracting entities.

In the third phase, the set of 17 notarial documents of the 19th century was tested, applying a combination of prompts from the first and second phase of experiments. The aim was to investigate the extent to which the used LLM (ChatGPT-3.5) is capable of understanding the complexity of the language by identifying a) the type of the contract, b) the contracting entities

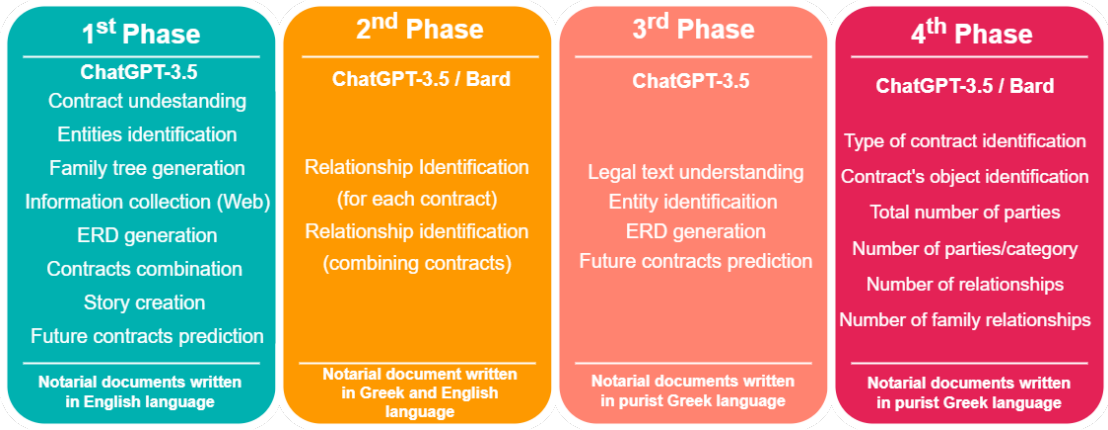

Figure 1: Summary of the research methodology outlined in 4 phases.

and their relationships, and c) to predict future legal acts between them.

The fourth and last experimental phase constitutes the main experiment, leveraging the experimental findings of the previous phases to craft the most appropriate prompts, while the same set of the 17 notarial documents and the two LLMs were used. Finally, the recorded responses of both LLMs were compared against the corresponding human responses. The examples and the experimental results are available in GitHub repo.[3]

4. Experiments

4.1. First phase

In the first experimental phase, three different examples of contracts were used (a pre-nuptial contract, a land concession contract, and a cooperation contract) written in English language. First, the pre-nuptial contract was used, while only the first 7 out of 33 pages of the document were inserted due to context window limitation of ChatGPT-3.5 to read such a large amount of information. Initially, the LLM was given five questions/prompts regarding: a) contract type understanding, b) entities identification, c) family tree generation, d) information from external sources (Web), and d) entity relationship diagram generation. ChatGPT-3.5 correctly identified both the type of contract and the involved entities while generating the corresponding family tree. Regarding the gathering of information about the individuals involved from external sources such as the Internet, ChatGPT-3.5 indicated a limitation related to internet connectivity, while also highlighting concerns about security and privacy. Finally, regarding the creation of entity relationship diagrams of the contracting entities, ChatGPT-3.5 presented a weakness in its construction.

The second goal was to investigate the ability of ChatGPT-3.5 to combine information from multiple contracts. To achieve this, the first two pages of the other two contracts were introduced to the LLM and three prompts were engineered regarding: a) combining information from

[3]https://github.com/AndreasSoularidis/LLM_historical_legal_documents

Table 1
Results of the first experimental phase

Research Questions	Success	Failure
Contract type understanding	✓	
Identifying entities	✓	
Creating family tree	✓	
Collecting information from external sources		✓
Creating diagram of entities and their relations		✓
Combining information from different contracts		✓
Creating history from different contracts	✓	
Predicting future legal acts	✓	

different contracts, b) crafting a narrative based on the given contracts, and c) forecasting future legal acts.

In more detail, first, ChatGPT-3.5 was prompted to combine information from the first two contracts creating a new one with persons listed in the first of them, in order to investigate its ability to combine information from different contracts. ChatGPT-3.5 indeed created a new contract regarding agricultural land lease using names of the persons of the first contract, also including elements written in both contract such as the rental fee and the signatures boxes at the end. However, when it was asked to create a new contract combining information from more than two contracts using the involved entities' names from one of them, this was accomplished only after a couple of unsuccessful attempts. Moreover, the generated contract did not contain the required names, and its content was poor. Next, ChatGPT-3.5 was asked and created a story combining all the given contracts. Finally, it was asked to predict future legal acts among the involved entities based on the generated story, to which it provided creative responses. The generated results are summarized in Table 1.

4.2. Second phase

In the second phase of experiments, a single type of contract (for simplicity and to facilitate thorough analysis) was used, specifically a declaration of inheritance acceptance. Initially, this contract was completed with data (names, surnames, etc.) in Greek, and then with data in English. Subsequently, the entire contract was translated into English to assess the efficacy of LLMs in both languages. At this stage, both ChatGPT-3.5 and Gemini/Bard were used to investigate to what extent they could identify the relationships between the involved entities, and whether they are collaborative relationships or kinship ties. In more detail, for the purpose of the study, the information of the contracting entities was initially inputted. In the first case, the father bequeaths his property to the son, and in the second case, a similar contract is filled out involving more entities, where the son of the first contract, now a father, bequeaths his own property to his children. Both LLMs were asked to identify the relationships of the involved entities a) for each contract separately and b) by combining the information from both contracts.

For the first question, the answers given by ChatGPT-3.5 were only a numbering of the attributes of these individuals (e.g., Manolis is the father who leaves his property to his son)

failing to identify the relationships between the entities, while for the second question it correctly identified the family relationships. On the other hand, Gemini/Bard identified kinship relations rather than property relations. When the contracts were filled in with data in English, ChatGPT-3.5 simply stated that the father leaves his property to his son, in both contracts, while failing to identify the grandfather grandchildren relationship. In contrast, Gemini/Bard answered all the questions correctly including the relationship of grandfather grandchildren. However, what is surprising is that Gemini/Baed didn't answer any questions when the entire contract was manually translated into English, except the last one where it correctly identified the family relationships. In contrast, ChatGPT-3.5 gave more accurate answers than it did in Greek but was still unable to understand the reference to the grandfather, failing to identify the relationship between the individuals of the given contracts. The results of the second experimental phase are summarized in Table 2.

Table 2
Comparison of Relationship Identification

Research Questions	Greek Text		English Names		English Text	
	ChatGPT	Bard	ChatGPT	Bard	ChatGPT	Bard
For each contract	X	✓	X	✓	✓	X
Combining contracts	✓	✓	X	✓	X	✓

4.3. Third phase

In the third phase, a combination of questions/prompts from the first and second phase of experiments were used, aiming to investigate the extent to which ChatGPT-3.5, and LLMs in general, are capable of understanding the intricacies of language. Particularly, the prompts concerned a) text understanding (regarding their ability to identify a notarial document), b) named-entity recognition c) generation of entity relationship diagram, and d) prediction of future legal acts between the involved entities. According to the final responses, ChatGPT-3.5 was capable of understanding the content of the contract, but often identified the text as a legal document written in ancient Greek rather than in the purist form of modern Greek. It was also able to identify entities such as persons, locations, time elements, etc.

Table 3
Results of the first experimental phase

Research Questions	Success	Failure
Text understanding	✓	
Entities identification	✓	
Creation of entity relationship diagram		✓
Prediction of future legal acts		✓

On the other hand, it presented a weakness in creating an entity relationship diagram but was able to generate text describing the relationships between the document's entities instead.

Finally, it was unable to predict legal actions between the involved entities, citing a limitation in forecasting future legal acts. The generated results of third experimental phase are summarized in Table 3.

4.4. Forth phase

At this stage, the final prompts were engineered based on the experimental results of the previous experiments, while both LLMs were used. Particularly, the selected prompts concerned a) type of contract, b) the contract's object, c) the total number of involved entities, d) the number of entities per category, e) the number of relationships among the entities, and f) the number of family relationships. Finally, the results were compared with the human researcher's responses, who performed the corresponding tasks manually, measuring precision and recall.

The experimental results show that ChatGPT-3.5 identified perfectly the type and the object of contract. However, regarding the latter, the total correct and the total possible correct answers (ambiguous answers) are counted towards succeeding more accurate presentation of the results. In particular, for the total documents, 76% of answers were correct while 24% were probably correct.

Regarding the total number of involved entities, whose number is between three and eight, in some cases ChatGPT-3.5 responded fewer people than it actually identified (e.g., it states that there are three persons, but it identifies four). Moreover, responses about the number of entities per category (e.g., witnesses, guarantors, etc.) show problems such as inability to identify all categories, all persons, double mentioned names, and mentions to the notary. Finally, in the last question about the number of related relationships, it should be noted that only three out of seventeen documents mention family relationships, while ChatGPT-3.5 successfully identified only one of them. The experimental results of the fourth experimental phase are summarized in Table 4, while diagrams are available in our repo on GitHub.

Table 4
Comparison of Precision and Recall for Different Research Questions

Research Questions	ChatGPT-3.5		Gemini/Bard	
	Precision	Recall	Precision	Recall
Number of involved entities	95%	76%	100%	74%
Number of involved entities per category	92%	71%	100%	74%
Number of entities' relationships	-	-	90%	84%
Number of related relationships	88%	88%	88%	88%

On the other hand, Gemini/Bard correctly answered all questions about the type and the object of contracts. As far as the total number of involved entities, Gemini/Bard in some cases there were inconsistencies between the identified names and the identified number of persons. In these cases, we counted as correct answer the ones that had identified all names involved. In addition, in a couple of examples Gemini/Bard included the notary even though it had been asked not to count it, presenting similar behavior with ChatGPT-3.5. Regarding the questions

about the number of involved entities per category and number of relationships among entities, Gemini/Bard's main characteristic in the first case is the grouping of persons into two more general categories (e.g., entities and witnesses) than those of the researcher. However, it is able to identify the entities regardless of the grouping it makes. Further, in the second case, Gemini/Bard understands the meaning of the word "relationship" without the need for repeated attempts to better understand the question. Finally, in the last question about the number of family relationships, Gemini/Bard identifies two out of three. However, in the last one, Gemini/Bard creates a hypothetical scenario where some persons could be related. Thus, it is considered more accurate to count one correct answer rather than two.

5. Discussion

The experimental results demonstrate that each LLM has its own peculiarities and limitations, especially when used for semantic analysis (e.g., contract type understanding, etc.) of historical Greek manuscripts (notarial deeds). ChatGPT-3.5 is highly capable of performing tasks such as named entity recognition, creating fictional stories using entities (people's names) from two contracts, and identifying the type and subject of a given contract. However, it presents a weakness in a number of tasks related to retrieving information from external documents, combining information from more than two legal documents, and predicting future legal acts. Furthermore, ChatGPT-3.5 shows a weakness in identifying the relationships between entities, especially in Greek contracts, while its performance is insufficient in terms of the total number of involved entities, failing to identify all categories and all entities per category. Finally, ChatGPT-3.5 failed to understand the concept of the number of relationships among entities, while it is worth mentioning its difficulty in understanding documents written in Greek.

On the other hand, Gemini/Bard exhibits strengths where ChatGPT-3.5 falls short. First, Gemini/Bard was capable to distinguish entity relationships, but its own weakness surprisingly appears in the English documents analysis, since it cannot return correct responses to any of the tested prompts except those concerning family relationships. Gemini/Bard presents a proficiency in identifying type and subject of contracts, achieving 100% accuracy, while demonstrates its dominance over ChatGPT-3.5 in terms of the total number of involved entities alone, and per category, achieving higher precision and recall. With regards to number of relationships among entities, Gemini/Bard is proved more capable than ChatGPT-3.5 to recognize the concept of "relationship" and without any need for additional human feedback. Furthermore, regarding the identification of family relationships, both LLMs identified one out of three relationships, although Gemini/Bard has identified one more relationship, this was a potential and not a true relationship. Finally, Gemini/Bard proves a remarkable capability in the semantic analysis of documents written in Greek.

However, there are a number of concerns that must be taken into account regarding the incorporation of LLMs into law. First and foremost, the use of AI models raises ethical considerations as these models are likely to have been exposed to (trained on) biased data. Last but not least, it is important to note the uncertainty inherent in LLM-generated responses, leading to inconsistent responses even for the same given prompt, which affects not only the reproducibility of the results but also the reliability of LLMs in general, especially in such a

crucial domain.

6. Conclusion

The experiments presented in this paper highlight LLM's general ability to understand, semantically analyze, and extract information from transcribed Greek notarial documents. However, their limitations in identifying relationships between entities require further investigation. Their difficulty in understanding the Greek language, and particularly the purist form of Greek language used in legal documents that has been used until the mid-20th century, constitute a challenge for the LLMs, as their responses in the experimental prompting demonstrate. Nevertheless, their accuracy measured with this preliminary experimentation may be considered encouraging for planning further future experiments. Future plans include the experimentation with more LLMs (GPT-4, Claude) to broaden the collection of information and results, creating a more comprehensive picture of their strengths and weaknesses. Moreover, the semantic analysis and comparison of contracts in multiple languages could be designed with the involvement of human experts familiar with the languages in question.

References

[1] B. Vlachou, D. P. Kasselouri, E. Kavakli, Exploitation and implementation of htr technology in greek handwritten archives, 5th Pan-Hellenic Conference on Digital Cultural Heritage-EuroMed 2023 5 (2024).

[2] D. Trautmann, A. Petrova, F. Schilder, Legal prompt engineering for multilingual legal judgement prediction, CoRR abs/2212.02199 (2022). doi:10.48550/ARXIV.2212.02199. arXiv:2212.02199.

[3] G. Muehlberger, L. Seaward, M. Terras, S. A. Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, et al., Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, Journal of documentation 75 (2019) 954–976.

[4] Z. Sun, A short survey of viewing large language models in legal aspect, CoRR abs/2303.09136 (2023). doi:10.48550/ARXIV.2303.09136. arXiv:2303.09136.

[5] A. Blair-Stanek, N. Holzenberger, B. V. Durme, Can GPT-3 perform statutory reasoning?, in: M. Grabmair, F. Andrade, P. Novais (Eds.), Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023, ACM, 2023, pp. 22–31. doi:10.1145/3594536.3595163.

[6] J. J. Nay, D. Karamardian, S. B. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. H. Choi, J. Kasai, Large language models as tax attorneys: A case study in legal capabilities emergence, CoRR abs/2306.07075 (2023). doi:10.48550/ARXIV.2306.07075. arXiv:2306.07075.

[7] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, J. Ge, Lawbench: Benchmarking legal knowledge of large language models, CoRR abs/2309.16289 (2023). doi:10.48550/ARXIV.2309.16289. arXiv:2309.16289.