Legal Knowledge and Information Systems
R. Markovich et al. (Eds.)

doi:10.3233/FAIA251593

Multilingual Legislative Definitions Retrieval and Generation Using LLM and Agentic AI

Leonardo ZILLI, Michele CORAZZA, Monica PALMIRANI and Salvatore SAPIENZA

CIRSFID - ALMA AI, University of Bologna, Italy

ORCiD ID: Leonardo Zilli https://orcid.org/0009-0007-4127-4875, Michele Corazza https://orcid.org/0000-0002-7288-6635, Monica Palmirani https://orcid.org/0000-0002-8557-8084, Salvatore Sapienza https://orcid.org/0000-0002-5429-5217

Abstract. The application of AI-based methods and Large Language Models (LLMs) to the legislative domain poses unique challenges, including the extensive usage of normative references, the complexity of legal language, as well as the ever-changing nature of legal documents. We propose a multilingual (English-Italian) LLM-based method to both retrieve and generate legislative definitions in the context of the European and Italian legislation. These definitions are a crucial aspect of legislative documents, as they create new meaning for specific concepts, and their generation is an open challenge for any automatic method. New definitions should not conflict with pre-existing ones, be consistent with the specific legal domain (e.g., food, energy, finance), and instead leverage them when necessary. Our method fosters a Retrieval Augmented Generation approach, using LLM and Agentic AI, which considers the validity of existing definitions, the hierarchy of legal sources, and investigates strategies to mitigate hallucinations in the generation of definitions. We provide a quantitative and qualitative evaluation of the results of our experiments.

Keywords. Definitions, Legislation, Large Language Models, Agentic AI, RAG

1. Introduction

Drafting legislative documents is a fundamental component of the legislative process and a pillar of the normative system. This complex task is carried out by specialized legislative experts, and it is demanding even for experienced professionals. In this context, the integration of Artificial Intelligence (AI) tools could offer valuable support across a wide range of drafting-related activities.

Among such tasks, identifying the into force correct legislative definitions is a crucial step in drafting laws. In legal context, "Definitions" typically follow a pattern that establishes a clear relationship between a term (*definiendum*) and its description (*definiens*). This structure often takes the form: "*definiendum* means *definiens*" (e.g., *"domain" means one or several data sets that cover specific topics;"*). The process of drafting definitions is composed of three main tasks.

1. Identifying the appropriate definiendum according to the domain of the act and avoiding ambiguities with other similar terminology already used, with different meaning, in other context (e.g., *"data"* in data protection law has a certain meaning, but it has a different one in the Data Act).
2. Retrieving existing legislative definitions of the same term, to prevent overlapping and conflicts between different descriptions.
3. If no legislative definition of the definiendum exists, drafting an appropriate definiens for the term, bearing in mind "neighbor" definitions in the same domain and the necessity of avoiding ambiguity.

Given such complexity (*identification-retrieval-drafting*), this paper aims to identify possible methods of increasing the quality and the efficacy of legislative definition drafting, in particular by proposing a combined pipeline which simplifies the retrieval and drafting stages, which present both a higher degree of complexity and less discretion on the drafter, thus being prone to the development of AI-assisting tools.

While the progressive spread of Large Language Models (LLMs) and Agentic AI in the legal domain [1] allows new possibilities in document retrieval and text generation, legal theory demands significant adaptations of such computational approaches to the domain of law [2]. In particular, it is worth considering that legal texts bear a significant complexity in the handling of temporal aspects: repealed, no longer in force, or amended definitions should not be used in the generation of a new definition that should be applicable today. Finally, some parts of the text such as legal references (e.g., "Regulation 2016/679") have a specific meaning which shall be properly handled and evaluated not only syntactically, but also from a temporal or spatial/jurisdictional perspective.

While acknowledging such challenges, this paper seeks answers to the following questions:

- **RQ1**: to what extent LLMs and Agentic AI can contribute to the retrieval of existing legislative definitions from legal corpora?
- **RQ2**: to what extent LLMs and Agentic AI can contribute to the generation of new legislative definitions?
- **RQ3**: to what extent LLMs and Agentic AI can manage peculiar elements of the legal domain, namely jurisdiction, temporal validity, and normative references?

2. Background

In the application of AI to the legislative domain, a multitude of peculiar challenges arise from theoretical and practical issues that are intrinsically part of any legislative document [3].The scope of this study regards legislative definitions. They can have a multitude of characteristics and can be classified according to the following types:

- **Definitions as constitutive rules** introduce new concepts, principles or institutions and constitute the "poetic part" of the law. They act as a shortcut between a term and its legal meaning (e.g., "X counts as Y in C"). The majority of the definitions used in this study belong to the family of constitutive rules
- **Definitions as exceptions** support legal reasoning by specifying exceptions or derogations. These function as sub-definitions that extend, reduce, or exclude cases based on conditions. Definitions may broaden a category (e.g., "cabin luggage"), narrow its scope (e.g., restricting "placing on the market"), or explicitly exclude cases (e.g., excluding air-cushion vehicles from the term "ship").

- **Definitions as memberships** preserve consistency in legal concepts while maintaining adaptability over time. They typically follow an "is-a" structure, integrating common-sense or emerging societal terms into existing legal categories.
- **Definitions as equivalences** establish a semantic equivalence between two terms, often using formulations such as "means an offence equivalent to". These definitions frequently refer to external normative sources, to which they anchor their definition.
- **Definitions as interpretations** provide clarifications, specifications, or technical explanations without creating new legal categories. For example, defining the "length of a vessel" based on physical measurement standards.
- **Definitions as technical descriptions** present purely technical information without additional legal function. An example includes defining "a mile" as exactly 1852 meters.

The different nature of definitions is not the only critical issue. The structured nature of legislative documents, which are composed of chapters, sections, articles, lists with points, etc., is semantically relevant for the interconnection between provisions that determines the first criterion of legal reasoning of the legal text (e.g., the position of the points in the list is meaningful). Another issue that arises when dealing with legislative documents is the usage of normative references, meaning portions of text that are used to invoke information from another document. Their usage is pervasive in legislative definitions, allowing a single definition to be used across multiple laws (e.g., "personal data" means personal data as defined in point 1 of Article 4 of Regulation (EU) 2016/679). Finally, one issue arises when, over time, legislative documents are amended, repealed, thus affecting also their definitions. This means that each document can exist in multiple (consolidated) versions that were in force at different times, with different text for the same term of definition.

To mitigate the complexity of these challenges, across our experiments we leverage the Akoma Ntoso XML standard [4,5], a standard for legal documents which supports a variety of legal documents, including legislative ones. Akoma Ntoso allows our approach to be time-aware and to deal with documents using point-in-time consolidated versions. The Akoma Ntoso standard also supports the markup of structural elements of the document (article, paragraph, etc), the annotation of normative references which contain the URI of the referenced document, and extensive support for various types of metadata. In particular, AKN includes a sophisticated methodology for the annotation of all the types of definitions, especially those that are split into many provisions but semantically connected (e.g., definition splits in two article, in different points in the list).

3. Related Work and motivation

The prevalence of Natural Language Processing tools and LLMs in various multidisciplinary fields is one of the main trends of the last years. In the legal domain, this trend is also prevalent [6], since the automatic analysis of legal text can unlock a multitude of benefits for the legal community. Legal text pose unique challenges, and even state-of-the-art models struggle to produce satisfactory results [7]. Nevertheless, like in other fields, the prevalence of transformer-based models [8] first and LLMs later [9,10] has contributed to the field of legal informatics. Lately, applications based on Retrieval Augmented Generation (RAG) [11] can mitigate some of the shortcomings of LLMs in the

legal domain by grounding the models with accurate information.

This aspect is crucial for models that handle legislative documents, as these enter into force, are amended and abrogated continuously, meaning that models would be out of date in a very short amount of time. Interestingly, one of the application that involve RAG for the augmentation of LLMs is one where it was used to ground a model for question answering in the legal domain [12,13]. In a similar fashion, LLMs have been used to identify and map concepts in legislation with the goal of verifying EU harmonisation [14]. In the context of AI applications that handle legislative definitions, one relevant contribution regards the creation of a suggestion tool based on the European EuroVoc thesaurus[1] to suggest definition inside the LEOS editor [15]. Another approach, which is based on a generative LLM, is LexDrafter, which uses a RAG-based framework for the generation of European legislative definitions [16], enhancing the prompt with contextual information.

4. Datasets

Our experiments were performed on three datasets (in Italian and English, given the legal knowledge of the research group and by the legal expert evaluators), belonging to two different jurisdictions::

- **Eur-Lex**: a set of European Regulations and Directives, which were downloaded in FORMEX format from the eur-lex portal and converted in Akoma Ntoso XML. Out of the complete dataset, which contains 14305 documents from 2010 to 2021, 889 have been annotated with definitions obtained using a symbolic AI approach. They are in English.
- **Italian legislation in Normattiva**: a set of 3196 Akoma Ntoso XML documents extracted from Normattiva, the portal containing Italian laws. Out of those, 401 contain annotated normative references, which were extracted using linguistic patterns in the documents. They are in Italian.
- **Chamber of Deputies Bill (PDL)**: a set of 3709 Italian bills from the 18th and 19th legislatures, extracted by us from the Italian Chamber of Deputies and converted to Akoma Ntoso XML. Out of those, 78 contain legislative definitions. They are in Italian.

Throughout these documents, we used simple rule-based systems to annotate the legislative definitions contained within them, using the formulas adopted for definitions in the respective legislative documents (among others, "{term} means ..." in EU documents, "per {term} si intende" for Italian documents). The annotation was integrated in the Akoma Ntoso format using the appropriate elements ("def" for the definendum, "defBody" for the definiens). The documents are stored in an eXist XML database to allow the system to extract data and metadata from the documents leveraging Akoma Ntoso.

5. Methods

The system uses the LangGraph[2] framework to produce a conversational agent which is able to perform two tasks:

1. The retrieval of existing legislative definition in the two jurisdictions;

[1] https://eur-lex.europa.eu/browse/eurovoc.html
[2] https://github.com/langchain-ai/langgraph

2. If no definition is found, the system generates a new one.

In order to perform these operations, we use an agent-based approach, where the model itself is able to select and use tools to achieve its goals. The first step in both pipelines is the extraction of a series of parameters, which are provided in plain text by the user and extracted by the model:

- **User query:** the entire text provided by the user to the model;
- **Definiendum:** the term that the user provides in the query, for which they are seeking a definition;
- **Legislation filter:** can take one of two values: "it" or "eu", which indicate that the user requested a retrieved or generated definition from the Italian or European jurisdiction, respectively. Additionally, the system can operate with no filter, in which case it will prioritize EU legislative definitions.
- **Date filter:** specifies temporal constraints applied by the user to the retrieval of definitions. This component is crucial for a system that operates in the legal domain, where norms are continuously amended and users might be interested in definitions as they existed in a specific point in time. Both open and close intervals can be used (e.g., "2021-10-23 - 2021-12-31"). Like the legislation filter, this parameter is optional.

The system then retrieves relevant definitions based on embeddings, specifically those produced by the BGE-M3 multilingual model [17]. This model enables the implementation of a hybrid retrieval approach, with a similarity function that incorporates a lexical component and a semantic one. In particular, the semantic (dense) similarity is obtained from the normalized hidden states of their encoder model when processing the "CLS" token. Given two embeddings e_q and e_r, then, the similarity is computed as their dot product (cosine similarity):

$$S_d = e_q \cdot e_r \tag{1}$$

BGE-M3 also produces sparse embeddings, using a ReLU activation function applied to the linear projection of the hidden state in correspondence of each word. Then, the similarity between two sentences is obtained from:

$$S_s = \sum_{t \in q \cap p} (w_{qt} \times w_{pt}) \tag{2}$$

Where p and q are two sentences, and w_{qt}, q_{pt} are the sparse embeddings of two words that belong to both sentences. Finally, we obtain the similarity between the user query and each definition as follows:

$$S = w_1 S_d + w_2 S_s \tag{3}$$

Where w_1, w_2 were set to 1.0 and 0.7 respectively, to prioritize semantic similarity over lexical similarity, as this is crucial to retrieve semantically similar definitions when the generation of a new definition is required. The multilingual nature of BGE-M3 allowed us to store the Italian and English vectors in the same vector database. From the previous step, we can extract the top 10 legislative definitions in terms of similarity to the user query. Our choice of 10 was motivated by a trade off between providing the model with relevant definitions and to avoid irrelevant ones, and we found 10 to be a good balance in

this context. Once the 10 definitions have been extracted, we filter out those that do not match the specified jurisdiction if required, meaning that we use definitions from EurLex for the "eu" filter, while we use Normattiva and PDL for the "it" one.

The next step in our pipeline is to build a timeline for each definition, containing each version and the range of dates for which it was in force. This is achieved through a query to the eXist XML database, which allows us to retrieve the successive consolidated versions of each document and all the definitions contained within them. Through the timeline of definitions, the system can then apply the temporal filter to the timelines and select only the temporally relevant definitions. The complete timeline is presented to the user, so they can assess which version of a given definition they should adopt. In order to provide the model with a broader context about definitions, we also incorporate metadata of EuroVoc terms associated with the original documents of each retrieved definition in the EurLex corpus as metadata in the Akoma Ntoso documents and stored along with the definitions inside the vector store. The next step in the pipeline is the choice, performed by the model, of which of the retrieved definitions present to the user. We prompt the LLM to adopt the user persona of a legal expert specialized in legislative drafting, with the goal to identify the definitions that should be provided to the user from a given list. In cases of ambiguity, the model is instructed to leverage the information from the EuroVoc terms extracted in the previous step. Furthermore, we instruct the model to prefer EU definitions in an attempt to favor the harmonisation of the EU law.

If the LLM does not find any definition to be relevant to the user query, it opts to generate a new one. In the generation prompt, the LLM is asked to adopt the user persona of a legal expert of the relevant jurisdiction, and the 10 most relevant definitions are provided as examples to leverage the model's in-context learning capabilities and mimic in terms of style, tone, structure and lexicon of existing definitions.

Our prompt for the generation of definitions was first drafted with the help of researchers from the legal domain. A small-scale qualitative evaluation of the generated EU legislative definitions was then performed using the procedure outlined in Section 8. This evaluation highlighted a series of issues with the results:

- **Technological neutrality:** the system should not produce definitions that mention specific companies or technologies;
- **Terminology**: in the EU, some terms have very specific meanings. For example, "Country" means a third party country, while "Member State" indicates one of the states that are members of the EU;
- **Invalid References:** if the system generates references, they should be valid and point to existing and in force legislative documents.
- **Usage of examples:** the generated definitions should be as general as possible, without relying on specific examples as the individual instances are subject to interpretation from the judges.

Thanks to this previous evaluation, we were able to modify the generation prompt to give precise instructions that could counteract some of the problematic tendencies of the model when generating definitions. The results of this effort are discussed in Sec. 8.

6. Experimental Settings

We evaluate the system using "Llama 3.3 70B Instruct" as the agent model, a relatively recent model from the LLama 3 family [18]. Crucially, it is an open weights and multilin-

gual model which is able to effectively operate as an agent and to provide instructions on how to use tools. The evaluation is carried out using four distinct datasets, each designed to target distinct functionalities of the system through both quantitative and qualitative analyses:

- **LexDrafter Dataset:** originally introduced in the evaluation of the LexDrafter framework [16], contains 1,330 legislative definitions related to 1,007 distinct English terms.
- **Italian Definitions Dataset:** contains 1478 Italian legislative definitions, randomly sampled from the 4,985 entries extracted from the Normattiva and PDL datasets.
- **EU Qualitative Dataset:** this dataset has been curated by legal experts and contains 73 English terms that are not found in any EU legislative documents.
- **IT Qualitative Dataset:** includes Italian equivalents of the 73 terms in the EU Qualitative Dataset, translated manually by legal experts.

6.1. Quantitative Generation Evaluation

A general assessment of the system's generation capabilities is conducted using the "LexDrafter Dataset" and the "Italian Definitions Dataset" to evaluate performances in the generation of English and Italian definitions, respectively. The generated definitions are evaluated using BLEU, ROUGE, BERTScore and BLEURT. For BERTScore, we used "distilbert-base-uncased" model for English definitions and "distilbert-base-multilingual-cased" for Italian ones. For BLEURT, we used the BLEURT-20 checkpoint, which supports both English and Italian. The system is prompted to define each term in the corresponding dataset language using the queries: *"What is the definition of* {*term*} *in the European Legislation?"* for English and *"Qual è la definizione di* {*term*} *nella legislazione italiana?"* for Italian. To prevent the system from retrieving existing definitions from its own corpus, we force their generation by excluding any definitions from the initial retrieval phase (Sec. 5) whose definiendum matches the requested term.

6.2. Qualitative Generation Evaluation

Since quantitative evaluations fail to evaluate crucial aspects such as definitions' legal soundness, internal consistency, and alignment with established legal frameworks, expert-based evaluations are necessary to assess the quality of the generated definitions in the context of the legal domain [19]. The qualitative evaluation of the system was conducted by a team of legal experts, who analyzed the definitions generated for the EU Qualitative Dataset and the IT Qualitative Dataset. Each definition was evaluated according to the following dimensions:

- **Accuracy**: Does the definition describe the correct legal content?
- **Contextual Appropriateness**: Is the definition appropriate for the specific legal domain and context?
- **Completeness**: Does the definition include all the essential elements necessary to fully describe the legal concept?
- **Consistency**: Is the definition internally coherent and free from contradictions?
- **Timeliness**: Does the definition contain up-to-date information?
- **References Accuracy**: Are the references used in the definition valid?

The experts evaluated each generated definition by assigning a score for each di-

mension using a five-point Likert scale, where 1 indicates the lowest level of adherence to the dimension and 5 the highest. In addition to these graded dimensions, two binary criteria were added to the evaluation, in order to detect the presence of critical issues:

- **Hallucination**: Does the definition contain any false or incorrect information, particularly in relation to the EU legislation?
- **Irrelevance**: Does the definition include information that is irrelevant to the specific legal term being defined?

7. Quantitative Evaluation

For the comparison between the Italian and European datasets in terms of quantitative metrics, it must be noted that while BLEU, ROUGE and BLEURT can in principle be compared between these two datasets, BertScore cannot, as it uses two different models for the two datasets. The first interesting finding when comparing the Italian and European dataset is the fact that the model seems to produce better scores for Italian than for the EU definitions for the comparable metrics. This might be due to the more formulaic nature of norms in the Italian legal tradition, which might produce longer standard sequences of words, which in turn improve the performance of our model.

In the comparison between our approach and the results obtained from LexDrafter using LLaMA-2 and Vicuna, we obtain the best scores for the BLEU and BLEURT metrics, while in terms of BERTScore our model produces comparable results for the Recall value and slightly lower Precision and F1. These results show that our approach is at the very least comparable with LexDrafter, albeit with a discrepancy when applying various metrics. While a small amount of deviation can happen when applying the same model across different library versions and architectures, the Precision and F1 differences should be outside of the margin of error for both BertScore and BLEURT, meaning that there is a significant difference within these metrics. The most probable explanation is that our definitions, while they seem to be more lexically aligned with the original ones, might be slightly more semantically divergent from the point of view of the distilBERT model, even if the BLEURT scores do not reflect this trend. The ROUGE scores for the LexDrafter models are not included in the original publication, so they were not included in the table.

Model	**BLEU**				**ROUGE**			**BERTScore**			**BLEURT**
	1	2	3	4	1	2	L	P	R	F1	
DefAgent-IT	0.34	0.26	0.22	0.19	0.35	0.21	0.31	0.85	0.87	0.86	0.51
DefAgent-EU	**0.26**	**0.16**	**0.12**	**0.09**	**0.35**	**0.19**	**0.31**	0.78	**0.81**	0.79	**0.49**
LD-Vicuna	0.25	0.13	0.07	0.04	–	–	–	**0.83**	**0.81**	**0.82**	0.47
LD-LLama-2	0.28	0.15	0.09	0.05	–	–	–	**0.83**	**0.81**	**0.82**	0.47

Table 1. Summary of the automated generation metrics scores for our models (DefAgent-IT and DefAgent-EU for Italian and European documents) and for the two models from LexDrafter (LD-Vicuna and LD-LLama-2).

8. Qualitative Evaluation

Metric	IT	EU
Accuracy	3.57	3.97
Contextual Appropriateness	4.32	4.28
Completeness	3.94	4.01
Consistency	4.23	4.19
Timeliness	4.06	4.41
References accuracy	2.25	2.16

Table 2. Average scores for each dimension

Metric	IT	EU
Hallucination Rate	0.10	0.08
Irrelevance Rate	0.10	0.04

Table 3. Hallucination and irrelevance rates

Table 2 shows the aggregate qualitative evaluation for each dimension as rated by the legal experts. Overall, the average level of accuracy is higher for the EU documents. This is not a surprise, given the higher number of training document in English in comparison to Italian. The generated definitions are usually appropriate in context and sufficiently complete, with both jurisdictions achieving satisfactory results in the context. They also maintain inner consistency, meaning that they are not contradictory. A significant gap can be observed in timeliness, with Italian definitions being less time-aware and point-in-time. Occasionally, while in scope and accurate, the legal reference provided for the Italian legislation has been repealed. Despite being instructed to evaluate solely in-force pieces of legislation, the system does not retrieve a valid reference.

High-valued definitions tend to be precise, contextually appropriate, and aligned with relevant legal and technological standards (when necessary), or the domain. An ideal definition clearly addresses the core elements of the concept without being limited to a specific sub-domain. For example, the definition of *Autonomous Trading Agent* demonstrates strong timeliness by accurately describing functionality without tying to tie the definition to a specific computational approach (e.g., Machine Learning). Similarly, the definition of *Gender-based Crime* performs well in terms of consistency, as it includes all relevant gender-related characteristics ("gender, gender identity, gender expression or sex characteristics") and refers to a specific legal act to ensure consistency with the legal framework. Similarly, in the references accuracy metric, well-drafted definitions cite the exact legal provision, such as a specific article of a regulation, thus ensuring legal clarity and traceability. Unfortunately, as the score show, this is not always the case. Most legal references are too broad (e.g., references to the Treaty of Functioning of the European Union), out of scope (e.g., references to other domains), not in force, hallucinated, redundant or recursive (e.g., using the definiens of the referenced law, which does not define the definiendum). Additionally, in terms of contextual appropriateness, certain definitions fail to reflect the practical implications or real-world application of the term, instead relying solely on abstract or overly legalistic formulations.

The results presented in Table 3 indicate an overall strong performance. Both the hallucination and irrelevance rates remain low across the IT and EU datasets. The hallucination rates of 0.10 (IT) and 0.08 (EU) suggest that the system generates mostly factually grounded content, while the irrelevance rates of 0.10 (IT) and particularly 0.04 (EU) highlight a sufficient degree of contextual relevance.

9. Discussion and Future Work

In this article we propose an agentic approach for the generation of legislative definitions in the context of European and Italian legislative documents. This approach leverages retrieved normative references from two datasets, then it decides whether to provide the user with a retrieved or generated definition for a given term. Furthermore, we allow the user to filter the results of their query using temporal and jurisdictional constraints. To the best of our knowledge, this is the first attempt to create a multilingual LLM-based approach for the generation of normative references, which is accompanied by a quantitative and qualitative evaluation of the results. In terms of the research questions discussed in the introduction, we can state that:

- **RQ1**: The retrieval of relevant definitions from a given corpora seems to be quite effective, and our agent is able to decide autonomously whether to generate or retrieve documents with a high level of accuracy.
- **RQ2**: Our agent-based approach is at least comparable existing method which is applied to the English language only. In terms of quantitative measures, achieving SOTA performance for the BLEU and BLEURT metrics.
- **RQ3**: Through our qualitative evaluation of the generated definitions, we understood clear limitations in the definition generation, namely with regards to the correct use of external legal references. Minor issues also concern the overall accuracy of the legal content.

In terms of future work, one of the main outcomes of the qualitative evaluation is the problematic nature of generated normative references within definitions. In this context, it would be useful to provide the model with some form of post-hoc evaluation of the references that it produces, allowing the LLM to correct its own output and to avoid the generation of references that are invalid, that point to amended or repeal portions of documents. This would require a careful implementation, as the point-in-time nature of the law means that references that are valid now might point to abrogated parts of documents in the future. This would also require the application of the approach to datasets that contain the entire collection of European and Italian legislative documents, as the validation of references requires access to all relevant legislative documents.

Another avenue might be a fine-tuning procedure that is specifically aimed at the generation of normative definitions. This would allow the model to learn the specific language used in legislative definitions, and avoid some of the issues that emerge with the contextual appropriateness of the results. However, this would not entirely solve the issues that arise from the generation of incorrect normative references.

The type of definitions that are used both for retrieval and for generation in our pipeline should also be expanded, as our model is only leveraging constitutive definitions, while other definitions (e.g., the interpretative ones) could also be included by keeping into account interpretations from the relevant courts in a given jurisdiction.

In conclusion, while the application of an agentic, RAG-based approach for the generation and retrieval of legislative definitions can be considered a successful experiments, the peculiarities and challenges that emerge when dealing with legislative documents is still partially unsolved, and further research is required on this task.

Acknowledgments

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185.

References

[1] Chau BK, Livermore MA. Computational legal studies comes of age. European Journal of Empirical Legal Studies, Virginia Public Law and Legal Theory Research Paper. 2024;(2024-40).

[2] Palmirani M. A smart legal order for the digital era: A hybrid ai and dialogic model. Ragion pratica. 2022;(2):633-55.

[3] Bresciani PF, Palmirani M. Constitutional opportunities and risks of AI in the Law-making process. Federalismi It. 2024;2:1-18.

[4] Palmirani M, Sperberg R, Vergottini G, Vitali F. Akoma Ntoso Version 1.0 Part 1: XML Vocabulary. OASIS Standard; 2018. Available from: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html.

[5] Vitali F, Palmirani M, Sperberg R, Parisse V. Akoma Ntoso Version 1.0. Part 2: Specifications. OASIS Standard; 2018. Available from: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html.

[6] Katz DM, Hartung D, Gerlach L, Jana A, II MJB. Natural Language Processing in the Legal Domain; 2023. Available from: https://arxiv.org/abs/2302.12039.

[7] Dahl M, Magesh V, Suzgun M, Ho DE. Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis. 2024;16(1):64-93.

[8] Greco CM, Tagarelli A. Bringing order into the realm of Transformer-based language models for artificial intelligence and law. Artificial Intelligence and Law. 2023:1-148.

[9] Siino M, Falco M, Croce D, Rosso P. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. IEEE Access. 2025;13:18253-76.

[10] Jayakumar T, Farooqui F, Farooqui L. Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. In: Preoțiuc-Pietro D, Goanta C, Chalkidis I, Barrett L, Spanakis G, Aletras N, editors. Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore: Association for Computational Linguistics; 2023. p. 223-9. Available from: https://aclanthology.org/2023.nllp-1.22/.

[11] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. .

[12] Improving Access to Justice with Legal Chatbots. Stats. 2020;3(3):356-75. Available from: https://www.mdpi.com/2571-905X/3/3/23.

[13] Mamalis ME, Kalampokis E, Fitsilis F, Theodorakopoulos G, Tarabanis K. A Large Language Model Agent Based Legal Assistant for Governance Applications. In: Janssen M, Crompvoets J, Gil-Garcia JR, Lee H, Lindgren I, Nikiforova A, et al., editors. Electronic Government. Cham: Springer Nature Switzerland; 2024. p. 286-301.

[14] Audrito D, Spada I, Mignone R, Sulis E, Di Caro L. Towards semi-automating European legislative harmonisation analysis: A harmonised glossary for LLM-based legal concept detection. Computer Law & Security Review. 2025;58:106171.

[15] Palmirani M, Vitali F, Longo G, Sante ED, Brega A, D'Arpa A, et al. Legal Drafting Supported by AI: Enhancing LEOS. In: Martino SD, Sansone C, Masciari E, Rossi S, Gravina M, editors. Proceedings of the Ital-IA Intelligenza Artificiale - Thematic Workshops. vol. 3762 of CEUR Workshop Proceedings. Naples, Italy: CEUR; 2024. p. 482-7.

[16] Chouhan A, Gertz M. LexDrafter: Terminology Drafting for Legislative Documents Using Retrieval Augmented Generation. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL; 2024. p. 10448-58. Available from: https://aclanthology.org/2024.lrec-main.913/.

[17] Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation; 2024. Available from: https://arxiv.org/abs/2402.03216.

[18] AI@Meta. Llama 3 Model Card; 2024. Available from: https://github.com/meta-llama/

llama3/blob/main/MODEL_CARD.md.

[19] Palmirani M, Sapienza S, Ashley K. A hybrid artificial intelligence methodology for legal analysis. BioLaw Journal-Rivista di BioDiritto. 2024;(3):389-409.