Published: 13 January 2026

Citation in BibTeX format

ICAIL 2025: 20th International Conference on Artificial Intelligence and Law
June 16 - 20, 2025
IL, Chicago, USA

RESEARCH-ARTICLE

Reporting Requests Modelling in European Legislation with a Hybrid AI Approach

MICHELE CORAZZA, University of Bologna, Bologna, BO, Italy

MONICA PALMIRANI, University of Bologna, Bologna, BO, Italy

SALVATORE SAPIENZA, University of Bologna, Bologna, BO, Italy

GENEROSO LONGO, University of Bologna, Bologna, BO, Italy

ICAIL '25: Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (June 2025)
https://doi.org/10.1145/3769126.3769255
ISBN: 9798400719394

Reporting Requests Modelling in European Legislation with a Hybrid AI Approach

Michele Corazza
CIRSFID ALMA-AI, Department of Legal Studies
University of Bologna
Bologna, Italy
michele.corazza@unibo.it

Monica Palmirani
CIRSFID ALMA-AI, Department of Legal Studies
University of Bologna
Bologna, Italy
monica.palmirani@unibo.it

Salvatore Sapienza
CIRSFID ALMA-AI, Department of Legal Studies
University of Bologna
Bologna, Italy
salvatore.sapienza@unibo.it

Generoso Longo
CIRSFID ALMA-AI, Department of Legal Studies
University of Bologna
Bologna, Italy
generoso.longo2@unibo.it

Abstract

Reporting requests in EU legislation require institutions to submit reports for legislative monitoring and policy implementation. As these meta-norms grow in complexity, compliance becomes increasingly challenging and burdensome. This paper leverages a Hybrid AI approach to detect, extract, and track reporting requests within the EU legislation using the RRVM ontology and European legal standards (AKN4EU, CELLAR, ELI). We investigate four interconnected and integrated areas: (1) detecting reporting requests and their concepts in legislative texts, (2) navigating normative references, (3) converting extracted legal knowledge into RDF for a Knowledge Graph, and (4) tracking modifications of reporting requests over time. Our approach compares machine learning (ML) and large language models (LLMs) for detection, demonstrating the advantages and limitations of both. By structuring Reporting Requests into a dynamic knowledge graph, our method improves the handling of Reporting Requests, reduces their administrative burden, and supports better legislative drafting and policy monitoring.

CCS Concepts

• **Applied computing** → **Law**; • **Computing methodologies** → **Natural language processing**; **Knowledge representation and reasoning**; **Information extraction**.

Keywords

Reporting Requests, Hybrid AI, Legislation, Ontologies, Knowledge Graphs

ACM Reference Format:
Michele Corazza, Monica Palmirani, Salvatore Sapienza, and Generoso Longo. 2025. Reporting Requests Modelling in European Legislation with a Hybrid AI Approach. In *20th International Conference on Artificial Intelligence and Law (ICAIL 2025), June 16–20, 2025, Chicago, IL, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3769126.3769255

ACM ISBN 979-8-4007-1939-4/25/06
https://doi.org/10.1145/3769126.3769255

1 Introduction and Research Questions

Reporting Requests (or requirements)[1] are particular legislative provisions that command some institutions to prepare a report, plan, budget, opinion, or standard to address other institutions for monitoring the quality of the legislation. Such commands are addressed to entities involved in the legislative system process to improve legislation quality, reduce burden, and implement European policies. Therefore, they are not norms in the classic meaning (addressed to the citizens) but meta-norms for managing and monitoring the law-making process. [2] The European Union (EU) legislation also includes extensive Reporting Requests on a diverse range of stakeholders, including public institutions, requiring the regular submission of information to other entities to allow such addresses to perform their tasks.[3]

Over recent years, the volume and complexity of these requirements have grown substantially, increasing the burden on EU institutions that often struggle to manage the overlapping mandates, especially with the evolution over time of the legislation due to amendments and derogations [26]. Failure to fulfil such mandates — whether through oversight, delays, or inaccuracies — can expose institutions to significant penalties and reputational risks, but also to not have a clear evolution of the needs to put in the legislative European Agenda for implementing the Better Regulation principles[4]. To mitigate these challenges, reporting requests are continuously monitored and updated by each EU Commission agency manually using SQL Databases. However, regulatory reporting remains a resource-intensive task that has led to calls for simplification

[1]Among their many definitions, "Reporting Requests" are occasionally referred to as "Regulatory Reporting"
[2]See the Case study analysis of regulatory reporting practices across the European Commission, https://op.europa.eu/en/publication-detail/-/publication/5a5e5b13-e996-11e9-9c4e-01aa75ed71a1
[3]See "The Importance of Metadata for Regulatory Reporting", Directorate-General for Informatics, D2 Interoperability, https://interoperable-europe.ec.europa.eu/sites/default/files/document/2021-12/Issue%20-%20The%20Importance%20of%20Metadata.pdf
[4]See the "Objectives of the Better Regulation agenda" at https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en

and burden reduction and the idea to automatically detect them during the legislative drafting and to monitor them over time are high-priority tasks[5].

For this reason, the European Commission has explored experimental approaches to alleviate the complexity of the tasks using legal ontologies, AI tools, legal drafting advanced support [6]. The Study on Regulatory Reporting Standards (SORTIS) project[7] exemplifies this effort, focusing on systematically identifying and formalising RRs within EU legislation. Such requests are structured to define the required actions, responsible institutions (e.g., the European Commission), and specified reporting temporal intervals using a formal ontology.

This paper aims at recognizing and extracting legal knowledge from RRs in European legislation using a hybrid AI approach, and to map this knowledge into an ontology called RRVM[8] which enables the tracking of RRs over time and to performing queries. This goal is meant to enable institutions to simplify the identification of RRs and monitor redundancies and overlapping while enabling advanced queries.

Our main aim is to build a legal knowledge graph using the European standards (AKN4EU[9], CELLAR[10], ELI[11], RRVM) to create a digital framework capable to detect, monitor, update the Reporting Requests and to permit sophisticated queries like:

"Give me all the Reporting Requests that the EU Commission shall produce, addressed to the European Parliament and Council (contextually), at biannual frequency, but suspended for COVID during the period 2019-2020 concerning standards about the quality of food."

"Give me all the Reporting Requests modified in the last 5 years, and the norms referred to as normative references in the consistent period of time according to the different consolidated versions that occurred".

To cope with these goals, we want to investigate the following research questions, which revolve around the extent to which we can:

- **RQ1**: Detect in the EU legislation AKN4EU document collection the reporting requests and extract the main semantic entities using the RRVM formal ontology as a reference
- **RQ2**: Navigate the normative references of EU legislation annotated in AKN4EU involved in the reporting requests, with a point-in-time methodology for extracting more information consistently with the period of time of the document.
- **RQ3**: Convert the legal knowledge entities extracted in RDF for populating a Knowledge Graph using RRVM as a reference for analysing the different kinds of reporting requests for future purposes like supporting legislative drafting with a template.
- **RQ4**: Use CELLAR RDF metadata to track the reporting requests over time for detecting possible modifications and so to easily update the Knowledge Graph and support Better Regulation principles (during the drafting or during the decision-making process of the policies).

This paper is divided into 8 sections. Following this introduction, Section 2 presents the related work and the motivation underlying this research. Then, Section 3 illustrates the Hybrid AI methodology followed in this study. While Section 4 presents the dataset, Section 5 describes the experimental settings. The results presented in Section 6 are then discussed in Section 7. Conclusive remarks and future perspectives are discussed in Section 8.

2 Background and Motivation

Classification of rules and extraction of elements from statutory texts are established fields of study in AI&Law [3]. The domain of automatic classification of legal norms is rich of examples in the literature, e.g., [10, 17, 33], as well as information extraction [30]. Machine Learning (ML) approaches have been used in health law classification with Decision Tress [19] and transfer learning [32]. In the field of extracting legally relevant elements from legal texts, studies have discussed the possibility of extracting legal rules on the basis of a semantic model for legislative texts by means of Support Vector Machine (SVM) algorithms [16]. More recent approaches have led to dedicated open-source packages to extract information from legal text built using Natural Language Processing (NLP) techniques and ML methods to offer features such as legal document segmentation, structured information extraction, named entity recognition (NER), and text conversion into feature vectors for the ML model [4]. As one study on legislation shows, legal word embeddings can enhance the performance of models by providing a more nuanced understanding of legal terminology and context. For instance, when classifying provisions into categories such as topics, embeddings trained on legal texts can help the model distinguish between terms that may have different meanings in legal contexts compared to general usage [8]. The same study discusses how embeddings enhance information extraction by improving the accuracy of tasks like NER by capturing domain-specific semantics. These embeddings allow models to recognize relationships between legal concepts.

Occasionally, classification and information extraction experiments have included ontologies or taxonomies aimed at supporting NLP algorithms in achieving a semantic-oriented approach. For instance, the literature has discussed the combination between and a NER tool to extract information from Spanish documents [5]. A more complex approach (Open Knowledge Extraction) has been proposed to leverage a privacy ontology [27] for identifying new potential classes following the identification of relevant information in privacy policies [25].

Transformer models and Large Language Models have significantly contributed to classification and extraction tasks, for instance, in the case of identifying the vague clauses in privacy policies [21]

[5]See, for instance, https://commission.europa.eu/system/files/2023-10/Factsheet_CWP_Burdens_10.pdf
[6]See the Streamlining regulatory reporting, https://interoperable-europe.ec.europa.eu/collection/better-legislation-smoother-implementation/streamlining-regulatory-reporting
[7]See the results available at https://interoperable-europe.ec.europa.eu/collection/better-legislation-smoother-implementation/news/streamlining-regulatory-reporting-sortis-project-results
[8]Available at https://interoperable-europe.ec.europa.eu/collection/better-legislation-smoother-implementation/solution/rrmv
[9]See AKN4EU: A Common Structured Format for EU Legislative Documents at https://op.europa.eu/en/web/eu-vocabularies/akn4eu
[10]See at https://op.europa.eu/en/web/cellar
[11]See at https://eur-lex.europa.eu/eli-register/about.html

or detecting unfair commercial practices [18]. Similarly, some experiments have shown potential in extracting information and classifying cases with LLMs [1, 6]. In the domain of the EU legislation, experiments have adopted an unsupervised semantic similarity approach supported by Transformer models to classify documents depending on their Sustainable Development Goals (SDG) targets [9].

Despite the promising results when compared to human-like evaluations [20], Transformer models and LLMs fall short of capturing the always-changing nature of the law, which happens every time the legislation is amended or declared unconstitutional [24].

Therefore, given the high priority of automating RR monitoring - also due to the financial burden for the EU Commission associated with this task - this paper is motivated by the necessity of developing an innovative methodological framework that can:

- keep into account modifications of statutory provisions in a diachronic and dynamic legal model which corresponds to the reality of the law rather than a flat and time-unaware representation of it
- support ML algorithms by means of a legal ontology designed to match the conceptual categories of the law rather than simply relying on semantics-unaware tokens
- compare conventional ML techniques and LLMs in classification and extraction tasks targeting statutory provisions while focusing on replicability to avoid the anchoring of the results to a given model.
- populate the ontology with the extract entities to maintain semantic consistency between the legal text, the ontology, and the algorithmic representation.

3 Methodology

The methodology adopted in this paper strictly derives from HAIMLA [28] and leverages a Hybrid AI approach. This section first clarifies what Reporting Requests are and the RRMV ontology, then presents the whole methodological framework before focusing on the annotation protocol.

3.1 Anatomy of a Reporting Request and the RRMV Ontology

First, legal research has been performed to analyse and formalise the elements of "Reporting Requests" (RRs) and the concepts correlated to them. We use the methodological approach proposed in Hybrid AI, for this reason the concepts surrounding the topic of this paper were formalised as a first step in a formal ontology. The experiments leveraged previous work. In particular, we used the RR Metadata Vocabulary (RRMV) ontology[12] designed to model RRs in the EU legislation.

The ontological definition of a `Request` is the following: "A `Request` is a command-to-do some `Action`, addressed to European institutions from the EU Commission, playing a given `Role` of addressee or addresser in a specific interval of time or with a frequency or in a given date. The `Request` produces `Action` to submit, produce, present, motivate, and describe in a given period of time. The `Action` has `Result` (e.g., report, plan, standard, motivated amendment proposal, etc.)". In our experiment, we confined the domain of Requests to informative documents explicitly defined as "Reports", "Plans", "Draft Technical standards", or informative documents that had an addresser, an addressee, and did not have a binding or a legislative nature.

[12]Available at https://interoperable-europe.ec.europa.eu/collection/better-legislation-smoother-implementation/solution/rrmv. The ontology was developed in the context of the SORTIS project (Study On Regulatory reporTIng Standards) coordinated by the Digital-ready Policymaking Team and the SEMIC Team of the Interoperability and Digital Government Unit of the European Commission, DIGIT.B2.

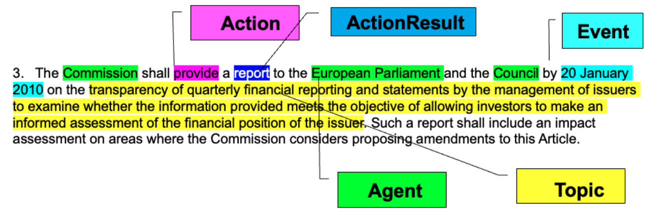

Figure 1: An example Reporting Request from Directive 2004/109/EC

Each `Request` may lead to an `Action`, which can be connected to other `Actions`. These connections may be semantic (via the `isRelatedTo` property) or sequential (using the `hasNext` property). `Action` can include annotations or statuses and involve an `AgentRole` (e.g., `Addreeseer` or `Addressee`), `ActionResult`, a specific `PeriodOfTime`, and a `Topic`. `Action` linked to a Request may have various statuses, such as completed, deleted, or overridden. Requests involve at least one `Agent` responsible for performing at least one `Action`. These, in turn, are events that produce at least one `ActionResult` within a defined `PeriodOfTime`[13].

The ontology was designed to annotate different patterns of RRs.

- including normative references relevant to the main components of the assertion:
  "Every three years, the Commission shall publish a summary based on the reports referred to in paragraph 4.";
- including multiple actions connected with different temporal parameters:
  "The Board of Supervisors shall adopt an annual report on the activities of the Authority, including on the performance of the Chairperson's duties, and shall, by 15 June each year, transmit that report to the European Parliament, to the Council, to the Commission, to the Court of Auditors and to the European Economic and Social Committee. The report shall be made public.";
- RR unspecified in the temporal event:
  "As soon as it adopts a delegated act, the Commission shall notify it simultaneously to the European Parliament and to the Council";
- RR conditioned:
  "By 31 December 2021, the Commission shall report to the European Parliament and to the Council on whether exceptional circumstances that trigger serious economic disturbance in the orderly functioning and integrity of financial more robust rule operators markets justify:(a)

[13]For more technical details of the RRMV ontology see here https://interoperable-europe.ec.europa.eu/collection/better-legislation-smoother-implementation/solution/rrmv/release/011

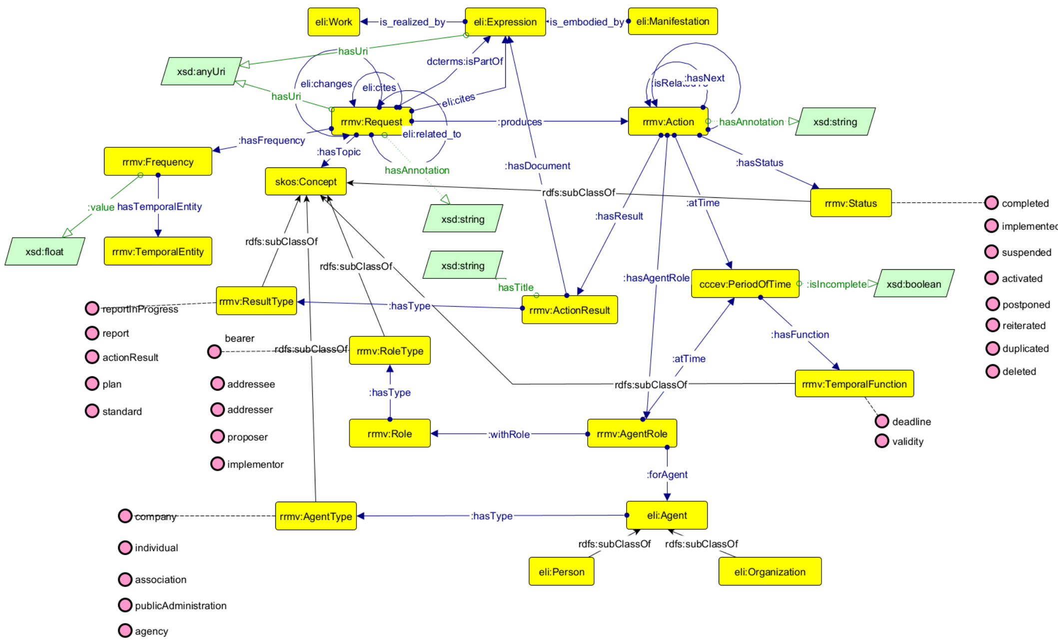

Figure 2: The RRVM Ontology

during such periods, permitting competent authorities to exclude from institutions' market risk internal models over shootings that do not result from deficiencies in those models;(b) during such periods, granting additional binding powers to competent authorities to impose restrictions on distributions by institutions.".

- RR split between different points in the nested lists:
  "By 26 June 2024, the Commission, in close cooperation with EBA and ESMA, shall submit a report, together with a legislative proposal if appropriate, to the European Parliament and to the Council, on the following:: (a), (b), (c) (i)(ii)(iii), (d), <omissis>, (h)"
  In this case, we intend to integrate RRMV with LegalRuleML https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/os/legalruleml-core-spec-v1.0-os.html ontology schema that offers robust deontic operators (e.g., permission, obligation, prohibition), defeasible logic structure (e.g., override, ranking with rules) and temporal reasoning framework.

Notably, the dataset used in this experiment does not include many records of these peculiar cases, and we definitely should improve the dataset in future work to obtain better balanced training on these specific structures.

3.2 Hybrid AI Framework

This study follows a systematic methodology for Hybrid AI legal text analysis, integrating both supervised ML and LLMs. Team members have mixed competencies (two experts in legal informatics and two computer scientists), but tasks are well-defined and separated in the implementation for avoiding to affect the neutrality of the annotation of the datasets. The methodology consists of six distinct stages visualised in Fig. 3:

(1) **Legal Analysis and Hypothesis Definition**
  - Formulation of the research question and hypothesis.
  - Conducting a preliminary legal analysis to establish the study's scope.

(2) **Preparation**
  - Identification of the relevant classes in the RRVM ontology to provide a structured legal categorization framework.
  - Compilation of a legal corpus.
  - Annotation of legal texts using the AKN XML Markup standard for semantic structuring and machine readability.

(3) **Annotation**
  - Manual annotation was divided into two steps:
    – binary classification of 991 provisions;
    – fine-grained annotation of 46 EU legislative AKN documents labelling the ontological entities within RRs.
    – Manual annotation of 493 concepts in 86 RRs according to the RRVM ontology.

(4) **Experiments**
  - Implementation of supervised ML and LLM classification models to analyze and categorize the annotated legal documents as "Reporting Requests" or not.
  - Implementation of supervised ML and LLM extraction models to populate the RRVM ontology.

(5) **Validation**
  - Performance measurement of the classification and extraction models.
  - Error analysis to assess model reliability and compare results

(6) **Modelling**

- Conversion of the results of the ML and LLM in RDF to populate the RRMV ontology
- Extraction of a Knowledge Graph from CELLAR to provide insights to legal operators on the lifecycle of the RR (e.g., amendments)

(7) **Output**
- Derivation of an answer to the research questions based on both model performance alongside legal insights.

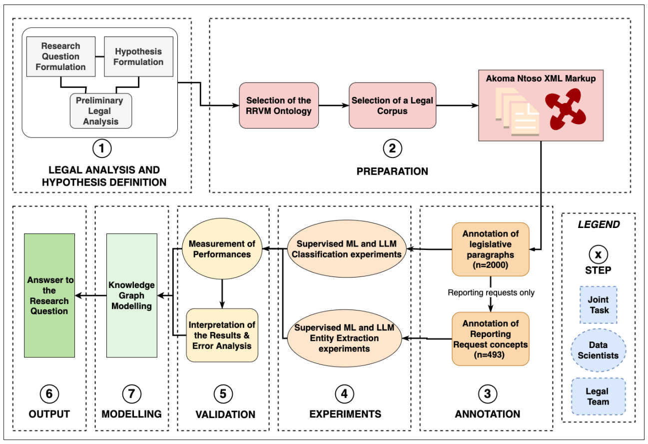

Figure 3: Hybrid AI Framework

3.3 Annotation protocol

The study comprises two complementary sets of experiments performed on two distinct datasets discussed below). For both datasets, the following annotation protocol was followed. Two legal experts (a senior and a junior) first agreed on the common definitions to adopt in the annotation. These definitions closely followed the descriptions of the classes in the RRMV ontology, both with regard to "RR" and the individual classes. For this experiment, an agreement was reached on annotating the following classes: `Addreeseer`, `Addressee`, `Action`, `ActionResult`, `PeriodOfTime`. The two experts have worked in parallel without contact with the rest of the team. First, the RR annotation was performed. Then, the same experts annotated the articles annotated as "RR" to identify the classes. The annotation was performed bearing in mind the following descriptions from the RRVM ontology.

```
<rdf:Description rdf:about="http://data.europa.eu/2qy/
     rrmv#addressee">
  <rdfs:comment xml:lang="en">Role played by an agent
       that is the destination of an action.</
       rdfs:comment>
</rdf:Description>

<rdf:Description rdf:about="http://data.europa.eu/2qy/
     rrmv#addresser">
  <rdfs:comment xml:lang="en">Role played by an agent
       that is producing an action.</rdfs:comment>
</rdf:Description>

<rdf:Description rdf:about="https://semiceu.github.io/
     CCCEV#PeriodOfTime">
  <rdfs:comment xml:lang="en">A temporal entity with non-
       zero extent or duration.</rdfs:comment>
</rdf:Description>

<rdf:Description rdf:about="http://data.europa.eu/2qy/
     rrmv#Action">
  <rdfs:comment xml:lang="en">An event with at least one
       agent that is participant in it in a given time.</
       rdfs:comment>
</rdf:Description>

<rdf:Description rdf:about="http://data.europa.eu/2qy/
     rrmv#ActionResult">
  <rdfs:comment xml:lang="en">The outcome of an Action.</
       rdfs:comment>
</rdf:Description>
```

4 Datasets

This study has used data from 48 EU legislative documents, from 1998 to 2021, including the consolidated text till 2023.

The Requirement classification dataset comprises 991 paragraphs for the binary classification and 86 paragraphs for the fine-grained annotation. The 86 paragraphs have been partitioned 66 for the training, 10 for the testing, and 10 for the validation for the qualitative analysis. We have additionally selected a further 50 EU legislative documents for validating the final results from legal point of view qualitative analysis).

The Entity extraction dataset comprises the full texts of the pieces of legislation containing RRs annotated as such by the legal experts in the first round of annotation. This data is an AKN format to allow the retrieval of valuable information, including legal references within the text. For instance, in the Requirement "EBA shall submit the draft technical standards referred to in paragraphs 4 and 5 to the Commission by 2014-12-31"[14], the proposed system should be able to navigate the legal reference ("*paragraphs 4 and 5*") to correctly identify the content of the Requirement and be useful to the user querying the system.

5 Experimental Settings

The automatic annotation of RRs can be divided into two different steps, which are performed using both a supervised ML approach and in-context learning with LLMs. In particular, the first step consists of detecting RRs within the legislative corpus, meaning that each sample should be categorized as either a RR or not. Then, the entities that are involved in each RR are extracted following the ontology described above: the `Addressee`, `Addresser`, `Action`, `ActionResult`, and `PeriodOfTime`. Once detected, these entities are used to create a knowledge graph that represents the requirements. All the experiments have been performed on machines with 4 Nvidia A100 GPUs. The classification tasks used a single GPU during training (and inference), while the inference of LLMs was performed on all 4 GPUs to allow the usage of non-quantized versions of the models. The training and inference procedures required less than 30 minutes for all models on the specified hardware. All the seeds were set to 42 for all experiments to allow their reproduction.

[14]Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC, art. 51(6)

Batch size	128
Epochs	100
Optimizer	AdamW
AdamW $\beta_1, \beta_2, \epsilon$	09, 0.999,1.0
Maximum gradient norm	1.0
Gradient Scheduler	linear
Patience (early stopping)	10
Learning Rate	10^{-5}

Table 1: Hyperparameters

5.1 Requirement classification

The first task to tackle the annotation of requirements is a binary classification: given a paragraph extracted from EU legislation, we want to know whether it contains a RR or not. The first step to achieve this goal is to create a train/validation/test split using a 60/20/20 ratio, which allows the evaluation of the performance across different approaches on the same test set.

Our first binary classification method uses pre-trained transformers, which are fine-tuned on the requirement classification task. In order to evaluate the impact of different models on the results, we selected 3 pre-trained models for our experiments:

- BERT base[11]: the original encoder model from Google AI was selected as a strong baseline for the other models in the experiments
- Legal BERT uncased EURLEX[7] is a BERT encoder model which was trained on data from EurLex, which is the official source for EU legislative documents
- ModernBERT[14] is a more recent encoder model, which uses the same base architecture as BERT but with multiple improvements such as rotary positional embeddings and unpadding. This model is characterized by a relatively large context window of 8K tokens, compared with the 512 of the original BERT.

All models were trained using the hyperparameters shown in Table 1 using the HuggingFace transformer library [34]. We adopted early stopping and used the F1 score of the minority class ("RR") on the validation set as a stopping criterion. Furthermore, we used class weights to adjust the cross-entropy loss to account for the imbalance in the dataset. The values chosen for the class weights were obtained from the scikit-learn library [29], which implements the approach described in [22].

The experiments are aimed at verifying whether LLMs could be used to detect and annotate RRs using in-context learning, meaning that they have access to just a few examples (few-shot learning) or even none (zero-shot learning). The experiments were conducted using LangChain, a framework to develop applications based on LLMs, which standardised the pipeline and the Hyperparameters across different open weights LLMs:

- Llama-3.3-70B-Instruct[2], which is the latest (to date) version of the Llama models from Meta and achieves similar performance to llama 3.1-400B for some tasks.
- Qwen2-72B-Instruct[13], a 72B parameter model from Alibaba, which has a similar size as Llama3.3 70B.
- Phi-3.5-MoE-instruct[15]: a small mixture of experts models from Microsoft, with 42B total parameters, of which only 6.6B are active during inference.
- gemma-2-27b-it[31]: a smaller-scale dense model from Google, trained with the same technology as Gemini.
- Mixtral-8x7B-Instruct-v0.1[12]: a mixture of experts models from Mistral, with a total of 46.7B parameters but which only uses 12.9B parameters during inference.

These models can be used as a good representation of the performance of small to medium-size LLMs, meaning that the results of these experiments should prove general even when this approach is applied to other LLMs.

The binary classification with LLMs was first conducted by providing only the description of a RR (zero-shot), 4 or 10 examples, respectively. We opted for an even number of examples in order to provide the LLM with an even number of positive and negative examples. An example of the prompt used for the models is shown below, where the squared brackets indicate the optional examples used during the few-shot procedure. We also omitted the format information provided by LangChain.

System: You are a classifier for REQUESTS in EU legislation. A REQUEST is a command to do some ACTION , addressed to European institutions from the EU Commission, playing a given ROLE of addressee or addresser in a specific interval of time or with a frequency or in a given date. The REQUEST produces ACTION to submit, produce, present, motivate, describe, in a given period of time. The ACTION has RESULT (e.g., report, plan, standard, motivated amendment proposal, etc.). You should determine whether a paragraph contains a REQUEST or not. Answer true if the paragraph contains a request, false otherwise.
[Examples]
Human: At least annually, the Authority shall consider whether it is appropriate to carry out Union-wide assessments of the resilience of financial institutions, in accordance with Article 32, and shall inform the European Parliament, the Council and the Commission of its reasoning. Where such Union-wide assessments are carried out and the Authority considers it appropriate to do so, it shall disclose the results for each participating financial institution.

5.2 Entity extraction

The second step in the automatic annotation of RRs consists of extracting entities associated with the request, i.e., `Addreseer`, `Addressee`, `Action`, `ActionResult`, `PeriodOfTime`. Similarly to the approach adopted for the binary classification task, the extraction of the entities has been conducted with using both a supervised ML approach and in-context learning with LLMs.

For the ML approach, the same models described in Section 5.1, with the hyperparameters shown in Table 1 were used. The model was trained to perform a token classification task, where each token belongs to no entity, or it is at the beginning or the middle of an entity. Due to the word piece tokenization approaches used by BERT and related models, we only applied labels to the beginning of words, ignoring the following tokens. We adopted entities using a "B-I" scheme, meaning that the beginning of an entity is annotated as "B", while all other tokens belonging to it are annotated as "I". As an example, we show the tokens and associated labels for one of the examples in our training set in Table 2.

The accuracy of the model on the entities (excluding "ignore" and "other") has been used for early stopping. During the reconstruction

Each	year	,	the	Exec	utive	Director	shall
B-periodOfTime	I-periodOfTime	O	O	B-addresser	ignore	I-addresser	O

Table 2: The tokens and their associated labels in a sample from the training set

of the entities, inconsistencies were fixed by applying the "B" class to tokens that are at the beginning of an entity but which the model predicts as "I". The subwords are also merged with the previous entity when necessary.

The second extraction experiment was conducted using LLMs in an in-context learning scenario. In particular, we apply a zero, 4 and 10 shot prompt to the LLM mentioned in Section 5.1 and ask the LLMs to extract a JSON object which contains all the entities present in a request. Additionally, the description of the relevant concepts from the RDF XML version of the ontology, namely the ones shown in Section 3.3, was used in a follow-up experiment. Crucially, these RDF descriptions were the same, which were used by the legal experts for the manual annotation of the corpus (reported above) consistently with the Hybrid AI approach. In particular, we conduct experiments with or without examples (zero or 4,10-shot) and with or without the RDF XML descriptions of the concepts. The prompt and its components are shown below, with the exception of the formatting instructions obtained from langchain, which are omitted.

System: You are an information extractor for REQUESTS in EU legislation. A REQUEST is a command to do some ACTION , addressed to European institutions from the EU Commission, playing a given ROLE of addressee or addresser in a specific interval of time or with a frequency or in a given date. The REQUEST produces ACTION to submit, produce, present, motivate, describe, in a given period of time. The ACTION has RESULT (e.g., report, plan, standard, motivated amendment proposal, etc.). Given a REQUEST, you should extract an addressee, an addresser, an ACTION, a RESULT, a period of time. **The description of these entities is specified in the following RDF:**
[RDF]
[Examples]
Human: 2. Following the review referred to in paragraph 1 , and after consulting ESMA, the Commission shall submit a report to the European Parliament and to the Council, accompanied, if appropriate, by a legislative proposal.

5.3 Semantic and Modelling

The results of the LLM are exported in JSON and serialized in RDF-XML using the RRMV, enabling the generation of a KG that dynamically integrates regulatory modifications.

```
<rrmv:Request rdf:about="http://data.europa.eu/2qy/rrmv#
    Request82">
  <dcterms:isPartOf rdf:datatype="http://www.w3.org
      /2001/XMLSchema#anyURI">/akn/eu/act/directive/
      elb/2001-03-12/18/eng@2018-03-29</
      dcterms:isPartOf>
  <rrmv:hasAnnotation>  5.    Every three years, the
      Commission shall publish a summary based on the
      reports referred to in paragraph 4.     </
      rrmv:hasAnnotation>
  <rrmv:hasUri rdf:datatype="http://www.w3.org/2001/
      XMLSchema#anyURI">/akn/eu/act/directive/elb
      /2001-03-12/18/eng@2018-03-29~
      sec_D__art_31__para_5</rrmv:hasUri>
  <rrmv:produces>
    <rrmv:Action rdf:about="http://data.europa.eu/2qy/
        rrmv#Request82Action1">
      <rrmv:atTime rdf:resource="http://data.europa.eu
          /2qy/rrmv#Request82Action1PeriodOfTime1"/>
      <rrmv:hasAgentRole rdf:resource="http://data.
          europa.eu/2qy/rrmv#
          Request82Action1AgentRole1"/>
      <rrmv:hasAgentRole rdf:resource="http://data.
          europa.eu/2qy/rrmv#
          Request82Action1AgentRole2"/>
      <rrmv:hasResult rdf:resource="http://data.europa.
          eu/2qy/rrmv#Request1Action1ActionResult1"/>
    </rrmv:Action>
  </rrmv:produces>
</rrmv:Request>
```

Additionally, we use the CELLAR triple-store for extracting the information about the modifications affecting the RR. The capability of extracting information and populating the ontology enhances compliance with RRs by providing a structured and up-to-date representation of evolving legal requirements, which makes the models suitable for real-world applications.

6 Results

In the following section, we discuss the numerical performance of our ML and LLM-based approaches for the extraction and annotation of RRs. The first step in a requirement annotation pipeline is a binary classification task, which we tackled with a multitude of different approaches. The performance of these models on the test set is shown in Table 3. By comparing the supervised models at the top of the table, we observe that every model reaches a good level of performance. In particular, the baseline BERT model is the worst-performing supervised model in terms of macro-averaged F1 score. We note a marked improvement in the detection of RRs when the more modern ModernBERT model is used. The specialized LegalBERT model proves to be the best-performing among the supervised ML approaches.

Model	Strategy	Other			RR			AVG		
		P	R	F1	P	R	F1	P	R	F1
BERT	supervised	0.987	0.952	0.969	0.516	0.800	0.627	0.751	**0.876**	0.798
LEGAL-BERT	supervised	0.978	**0.990**	**0.984**	**0.813**	0.650	**0.722**	**0.895**	0.820	**0.853**
ModernBERT	supervised	0.984	0.977	0.981	0.682	0.750	0.714	0.833	0.864	0.847
Llama 3.3	zero-shot	0.995	0.626	0.768	0.141	0.950	0.245	0.568	0.788	0.507
	4-shot	0.995	0.613	0.758	0.137	0.950	0.239	0.566	0.781	0.499
	10-shot	0.991	0.713	0.829	0.168	0.900	0.283	0.580	0.806	0.556
Mixtral-8x7B	zero-shot	0.976	0.803	0.881	0.187	0.700	0.295	0.582	0.752	0.588
	4-shot	0.971	0.855	0.909	0.211	0.600	0.312	0.591	0.727	0.610
	10-shot	0.974	0.955	0.964	0.462	0.600	0.522	0.718	0.777	0.743
Gemma 2 27B	zero-shot	**1.000**	0.306	0.469	0.085	**1.000**	0.157	0.543	0.653	0.313
	4-shot	**1.000**	0.594	0.745	0.137	**1.000**	0.241	0.568	0.797	0.493
	10-shot	0.996	0.777	0.873	0.216	0.950	0.352	0.606	0.864	0.613
Phi-3.5-MoE	zero-shot	0.990	0.610	0.754	0.129	0.900	0.226	0.560	0.755	0.490
	4-shot	**1.000**	0.442	0.613	0.104	**1.000**	0.188	0.552	0.721	0.400
	10-shot	0.992	0.761	0.861	0.196	0.900	0.321	0.594	0.831	0.591

Table 3: Comparison of Metrics Across Classifiers. The highest overall value for each metric is indicated in bold. The underline indicates the highest value of each metric for supervised models and LLMs, respectively

The second set of experiments involves the application of a few-shot scenario for the classification of RRs. In this case, the LLMs struggled to produce good-quality results for the RR class. In particular, they all produced a high number of false positives, as evidenced by the high recall value. This effect might be an artefact of our approach, which includes 50% of the samples from the RR class for few-shot learning. Among the LLMs, Mixtral 8x7B produces the best overall result, which is characterized by a higher F1 value for the minority class. The difference between the 10-shot Mixtral model and the other ones is quite severe, but Llama, Gemma and Phi all perform similarly on this task. In most cases the usage of examples from the training set produces higher quality results, with the larger difference between 4 and 10-shot. Overall, supervised models proved more effective in the detection of RR, as they were able to leverage the information contained in the training data. Conversely, the LLMs had very poor results for the annotation of requests, as they were producing a large number of false positives.

The next step of our approach is the extraction of entities with are associated with RRs. In order to evaluate the results of our approaches, we opted to use a strict entity-based evaluation similar to those adopted for NER tasks. Namely, we counted an entity as correctly identified (a true positive) only when it was completely identical to the one annotated in the gold standard. This choice was motivated by the nature of legal documents. As an example, when an EU legislative document mentions "the Council", it refers to the Council of the EU, composed of the chiefs of government of the Member States. If a RR was to be addressed to the "European Council", an entity extraction model might detect "Council" as the addressee, a completely different entity. The results of our evaluation can be found in Table 4.

In terms of performance, the supervised classification models are once again the best performing ones in this experiment. This time, ModernBERT is the worst performing model among the three, while BERT has a very similar performance to LEGAL-BERT, which shows the best overall performance in terms of macro averaged F1 score. Interestingly, however, all the supervised models show very poor performance on the PeriodOfTime class.

When observing the performance of LLMs, we can observe that, overall, they are not able to match the performance of the supervised models. In terms of overall trends, the usage of at least some examples tends to be beneficial, with the higher improvements between zero and 4-shot. Conversely, the usage of the RDF XML definitions inside the prompt does not seem to consistently improve performance, even though the second best result in terms of macro averaged F1 is obtained from Llama 3.3 using 4-shot and the RDF information. Interestingly, the number of parameters of each model does not correlate with their performance since the best results are obtained from 10-shot Gemma, followed closely by the larger dense model. Furthermore, our two "mixture of experts" models seem to achieve limited performances with this task, with Mixtral in particular struggling to extract action results.

Overall, while the supervised model had the highest performance scores, the LLMs seem more promising in the extraction of entities, as they were able to extract some information even from the challenging PeriodOfTime class.

7 Discussion

When comparing the supervised method, we observed that LEGAL-BERT produces higher quality results than ModernBERT and BERT in both experiments. While this discrepancy might be due to a multitude of factor, the fact that LEGAL-BERT has been trained on EurLex data is probably the main reason for the observed performance, meaning that the model is familiar with the language used in EU legislative documents.

The comparison of ModernBERT and BERT is less clear-cut, as ModernBERT shows a level of performance compared to LEGAL-BERT for the binary classification task, while BERT is close to LEGAL-BERT in the entity extraction. A more thorough evaluation of the models might highlight the reason behind these discrepancies.

In terms of overall trends, we can observe that the models seem to extract `Action` and `Addresser` better than other entities. Since most RRs are sent to very prominent EU institutions (typically the Commission, EU Parliament, etc.), they are more easily identifiable than the addresser, which might involve less prominent institutions. One of the more challenging categories is also `ActionResult`, perhaps due to the fact that this category is characterized by relatively long sequences of text, which are harder to segment accurately. Finally, one interesting aspect is the fact that the supervised models fail to identify correctly the temporal constraints which are applied to the reporting requests, with only a minor level of success with some LLMs. This result might be caused by a large variety in the types of temporal constraints and to the complexity of the temporal model in the legal domain, which might include dates as deadlines of the RR but also dates for references, content parameters (e.g., dates of payment). Additionally the linguistic variants involve frequency, dates, temporal spans, time algebra (after, before, between, end, beginning), but also exceptions that need a more expressive ontology (e.g., LegalRuleML).

Finally, it is relevant to stress that our results guarantee explainability, replicability, and sustainability. By linking the extracted information to a legal ontology which corresponds to the conceptual categories of the EU institution both in training and output, users can easily monitor the adherence of the extracted information to their conceptual classes. The ontological constraint also contributes to mitigate LLMs' "hallucinations", which a well-known issue in the legal domain[23]. Moreover, by introducing fixed seed values, the training is replicable with the code provided[15]. Finally, the training of the models has taken place in a High Performance Computing facility with strict environmental policies and ISO 50001:2018 certification to minimise carbon footprint and emissions during the training.

8 Conclusive Remarks

This paper answers to the following research questions:

- **RQ1**: By using different methods of LLM and classic ML (supervised), we have detected RRs with the best F1 of 0.863, with supervised ML approaches consistently outperforming LLMs.
- **RQ2**: By using the AKN4EU markup and the naming convention based on FRBR we have navigated the normative references and included the text in the model.

[15]The source code, datasets, RDF, JSON, AKN4EU, additional metrics, and the output of annotation are available at https://gitlab.com/CIRSFID/eulegislation-reporting-requests

Model	Strategy	Action			Addressee			Addresser			PeriodOfTime			ActionResult			AVG		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT		0.467	0.500	0.483	0.583	0.667	0.622	0.818	0.692	**0.750**	0.045	0.062	0.053	0.556	0.625	0.588	0.494	0.509	0.499
LEGAL-BERT	supervised	0.333	0.429	0.375	0.762	**0.762**	**0.762**	0.750	0.692	0.720	0.059	0.062	0.061	0.625	0.625	0.625	**0.506**	**0.514**	**0.509**
ModernBERT		0.600	0.429	0.500	0.684	0.619	0.650	0.571	0.308	0.400	0.000	0.000	0.000	0.364	0.500	0.421	0.444	0.371	0.394
Llama 3.3	zero-shot	0.500	0.143	0.222	0.667	0.571	0.615	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.233	0.143	0.168
Llama 3.3	zero-shot rdf	0.500	0.143	0.222	0.706	0.571	0.632	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.241	0.143	0.171
Llama 3.3	4-shot	**1.000**	**0.714**	**0.833**	0.667	0.095	0.167	0.450	0.692	0.545	0.167	0.062	0.091	0.667	0.250	0.364	0.590	0.363	0.400
Llama 3.3	4-shot rdf	**1.000**	**0.714**	**0.833**	0.667	0.095	0.167	0.474	0.692	0.562	0.200	0.062	0.095	**1.000**	**0.500**	**0.667**	0.668	0.413	0.465
Llama 3.3	10-shot	**1.000**	**0.714**	**0.833**	0.667	0.190	0.296	0.550	**0.846**	0.667	0.200	**0.125**	0.154	0.000	0.000	0.000	0.483	0.375	0.390
Llama 3.3	10-shot rdf	**1.000**	**0.714**	**0.833**	0.600	0.143	0.231	0.550	**0.846**	0.667	0.200	**0.125**	0.154	0.000	0.000	0.000	0.470	0.366	0.377
Mixtral-8x7B	zero-shot	0.571	0.286	0.381	0.471	0.381	0.421	0.333	0.077	0.125	0.000	0.000	0.000	0.000	0.000	0.000	0.275	0.149	0.185
Mixtral-8x7B	zero-shot rdf	0.625	0.357	0.455	0.611	0.524	0.564	0.333	0.077	0.125	0.000	0.000	0.000	0.000	0.000	0.000	0.314	0.192	0.229
Mixtral-8x7B	4-shot	0.900	0.643	0.750	0.000	0.000	0.000	0.444	0.308	0.364	0.000	0.000	0.000	0.000	0.000	0.000	0.269	0.190	0.223
Mixtral-8x7B	4-shot rdf	**1.000**	0.500	0.667	0.000	0.000	0.000	0.400	0.308	0.348	0.000	0.000	0.000	0.000	0.000	0.000	0.280	0.162	0.203
Mixtral-8x7B	10-shot	**1.000**	0.643	0.783	0.000	0.000	0.000	0.556	0.385	0.455	0.143	0.062	0.087	0.000	0.000	0.000	0.340	0.218	0.265
Mixtral-8x7B	10-shot rdf	**1.000**	0.571	0.727	0.000	0.000	0.000	0.556	0.385	0.455	0.143	0.062	0.087	0.000	0.000	0.000	0.340	0.204	0.254
Gemma 2 27B	zero-shot	**1.000**	0.214	0.353	0.737	0.667	0.700	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.347	0.176	0.211
Gemma 2 27B	zero-shot rdf	**1.000**	0.214	0.353	0.778	0.667	0.718	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.356	0.176	0.214
Gemma 2 27B	4-shot	**1.000**	**0.714**	**0.833**	0.750	0.143	0.240	0.444	0.615	0.516	0.143	0.062	0.087	**1.000**	0.250	0.400	0.667	0.357	0.415
Gemma 2 27B	4-shot rdf	**1.000**	**0.714**	**0.833**	0.750	0.143	0.240	0.389	0.538	0.452	0.200	0.062	0.095	**1.000**	0.375	0.545	0.668	0.367	0.433
Gemma 2 27B	10-shot	**1.000**	**0.714**	**0.833**	**0.800**	0.190	0.308	0.500	0.615	0.552	**0.400**	**0.125**	**0.190**	0.750	0.375	0.500	**0.690**	0.404	0.477
Gemma 2 27B	10-shot rdf	**1.000**	**0.714**	**0.833**	0.667	0.095	0.167	0.533	0.615	0.571	0.286	**0.125**	0.174	0.750	0.375	0.500	0.647	0.385	0.449
Phi-3.5-MoE	zero-shot	0.400	0.286	0.333	0.737	0.667	0.700	0.333	0.077	0.125	0.000	0.000	0.000	0.000	0.000	0.000	0.294	0.206	0.232
Phi-3.5-MoE	zero-shot rdf	0.571	0.286	0.381	0.786	0.524	0.629	**1.000**	0.077	0.143	0.000	0.000	0.000	0.000	0.000	0.000	0.471	0.177	0.230
Phi-3.5-MoE	4-shot	**1.000**	**0.714**	**0.833**	0.000	0.000	0.000	0.417	0.385	0.400	0.000	0.000	0.000	0.500	0.375	0.429	0.383	0.295	0.332
Phi-3.5-MoE	4-shot rdf	**1.000**	**0.714**	**0.833**	0.000	0.000	0.000	0.417	0.385	0.400	0.000	0.000	0.000	0.500	0.375	0.429	0.383	0.295	0.332
Phi-3.5-MoE	10-shot	0.909	**0.714**	0.800	0.000	0.000	0.000	0.385	0.385	0.385	0.000	0.000	0.000	0.500	0.125	0.200	0.359	0.245	0.277
Phi-3.5-MoE	10-shot rdf	0.909	**0.714**	0.800	0.000	0.000	0.000	0.385	0.385	0.385	0.000	0.000	0.000	0.500	0.125	0.200	0.359	0.245	0.277

Table 4: Performance metrics for the entity extraction in RRs. The highest overall value for each metric is indicated in bold. The underline indicates the highest value of each metric for supervised models and LLMs, respectively.

- **RQ3**: We have converted the result of the LLM or ML in RDF for populating the RRMV ontology.
- **RQ4**: We have extracted from CELLAR[16] a KG for tracking the lifecycle of each document, including RRs to offer to the legal expert the history of the RR. This allows the legal expert to verify whether some RRs are reiterated over time (e.g., due to economic crises, pandemics like COVID-19, failure to implement the expected policy) and monitor the overall effectiveness of the RRs over time.

The analysis of the results tells us that LLMs are more promising then ML (e.g., LEGAL-BERT) for entity extraction, as they are able to recognize even more challenging entities. Some classes of the RR are very difficult to detect even with LLM (`PeriodOfTime`), while others (`Addresser`, `Action`) are very easy to model, and different LLMs produce different F1 scores. Withini LLMs, Gemma produces the best average F1 score for this task.

Additionally, to these results, we have analysed the different outputs from a legal point of view, and this produced new inputs for the RRMV ontology. For instance, we have detected more complex temporal events (e.g., "No sooner than 26 June 2028", "At the earliest possible opportunity after 13 February 2021", "No later than 12 January 2032") and information about the execution of the tasks required (e.g. *ex-post* evaluation of the effectiveness of the RR), under which it is possible to execute the RR, or it is suspended. The report of the European Parliament on the "Review clauses in EU legislation adopted during the first half of the ninth parliamentary term (2019-2024)"[17] states that the 60% of the EU legislation includes RRs, with a more frequent timeframe of 1 year, repeated for 4 years. For this reason, we would like to include some rules coming from data analytics (statistical figures) for better detecting the other elements in the text.

On the basis of these results, we have future goals:

- to improve the dataset for balancing the odd cases;
- to annotate also the interconnection between `Request`, `Actions`, `PeriodOfTime`;
- to model conditionals in LegalRuleML using the information extracted;
- to design template-patterns for legal drafting tools (e.g., [18]);
- to involve EU Commission DG2 in the real evaluation of the tool;

[16] https://eur-lex.europa.eu/eli-register/technical_information.html

[17] Available at https://www.europarl.europa.eu/RegData/etudes/STUD/2022/734675/EPRS_STU(2022)734675_EN.pdf

[18] https://interoperable-europe.ec.europa.eu/sites/default/files/document/2022-06/Drafting%20legislation%20in%20the%20era%20of%20AI%20and%20digitisation%20%E2%80%93%20study.pdf

- to visualise the history of an RR over time for the decision-maker.

In the future, the RR modelled with RRMV can help legal drafters to better write, monitor, and evaluate the effect of the RR, which is an important instrument for implementing the Better Regulation principles in EU legislation. We have demonstrated that the Hybrid AI (LLM + Ontology RRMV + AKN4EU) can in any case sensitively reduce the burden of legal experts in managing this so relevant part of the EU legislation, thus contributing to improving the effectiveness of EU policies.

Acknowledgments

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185. The authors acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

References

[1] Subinay Adhikary, Procheta Sen, Dwaipayan Roy, and Kripabandhu Ghosh. 2024. A case study for automated attribute extraction from legal documents using large language models. *Artificial Intelligence and Law* (2024), 1–22.
[2] AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
[3] Kevin D Ashley. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age.* Cambridge University Press.
[4] Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2021. LexNLP: Natural language processing and information extraction for legal and regulatory texts. In *Research handbook on big data law.* Edward Elgar Publishing, 216–227.
[5] María G. Buey, Angel Luis Garrido, Carlos Bobed, and Sergio Ilarri. 2016. The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence* (Rome, Italy) *(ICAART 2016)*. SCITEPRESS - Science and Technology Publications, Lda, Setubal, PRT, 438–445. https://doi.org/10.5220/0005757204380445
[6] Chaitra C R, Sankalp Kulkarni, Sai Rama Akash Varma Sagi, Shashank Pandey, Rohit Yalavarthy, Dipanjan Chakraborty, and Prajna Devi Upadhyay. 2024. LeGen: Complex Information Extraction from Legal sentences using Generative Models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis (Eds.). Association for Computational Linguistics, Miami, FL, USA, 1–17. https://doi.org/10.18653/v1/2024.nllp-1.1
[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261
[8] Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* 27, 2 (2019), 171–198.
[9] Michele Corazza, Monica Palmirani, Franco M. T. Gatti, and Salvatore Sapienza. 2024. Monitoring Sustainable Development Goals in European Legislation using Hybrid AI. In *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance (ICEGOV '24)*. Association for Computing Machinery, New York, NY, USA, 261–269. https://doi.org/10.1145/3680127.3680223
[10] Emile de Maat and Radboud Winkels. 2008. Automatic classification of sentences in dutch laws. In *Legal Knowledge and Information Systems.* IOS Press, 207–216.
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423
[12] Albert Q. Jiang et al. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG] https://arxiv.org/abs/2401.04088
[13] An Yang et al. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] https://arxiv.org/abs/2407.10671
[14] Benjamin Warner et al. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663 [cs.CL] https://arxiv.org/abs/2412.13663
[15] Marah Abdin et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219
[16] Enrico Francesconi. 2010. Legal rules learning based on a semantic model for legislation. In *Workshop Programme.* 46.
[17] Enrico Francesconi and Andrea Passerini. 2007. Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law* 15 (2007), 1–17.
[18] Andrea Galassi, Francesca Lagioia, Agnieszka Jabłonowska, and Marco Lippi. 2024. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law* (2024), 1–49.
[19] Matthias Grabmair, Kevin D Ashley, Rebecca Hwa, and Patricia M Sweeney. 2011. Toward extracting information from public health statutes using text classification machine learning. In *Legal Knowledge and Information Systems.* IOS Press, 73–82.
[20] Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2023. Can GPT Alleviate the Burden of Annotation? In *Legal Knowledge and Information Systems.* IOS Press, 157–166.
[21] Giulia Grundler, Rūta Liepiņa, Mariaceleste Musicco, Francesca Lagioia, Andrea Galassi, Giovanni Sartor, and Paolo Torroni. 2024. Detecting Vague Clauses in Privacy Policies: The Analysis of Data Categories Using BERT Models and LLMs. In *Legal Knowledge and Information Systems.* IOS Press, 72–83.
[22] Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis* 9, 2 (2001), 137–163.
[23] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv:2405.20362 [cs.CL] https://arxiv.org/abs/2405.20362
[24] Monica Palmirani. 2022. A Smart Legal Order for the Digital Era: A Hybrid AI and Dialogic Model. *Ragion pratica, Rivista semestrale* 2/2022 (2022), 633–655. https://doi.org/10.1415/105387
[25] Monica Palmirani, Giorgia Bincoletto, Valentina Leone, Salvatore Sapienza, and Francesco Sovrano. 2020. Hybrid Refining Approach of PrOnto Ontology. In *Electronic Government and the Information Systems Perspective*, Andrea Kő, Enrico Francesconi, Gabriele Kotsis, A Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 3–17.
[26] Monica Palmirani and Davide Liga. 2022. Derogations Analysis of European Legislation Through Hybrid AI Approach. In *Electronic Government and the Information Systems Perspective*, Andrea Kő, Enrico Francesconi, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 123–137.
[27] Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. 2018. Legal ontology for modelling GDPR concepts and norms. In *Legal Knowledge and Information Systems.* IOS Press, 91–100.
[28] Monica Palmirani, Salvatore Sapienza, and Kevin Ashley. 2024. A hybrid artificial intelligence methodology for legal analysis. *BioLaw Journal-Rivista di BioDiritto* 3 (2024), 389–409.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[30] Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, and Vittoria Pistone. 2024. DiscoLQA: zero-shot discourse-based legal question answering on European Legislation. *Artificial Intelligence and Law* (2024), 1–37.
[31] Gemma Team. 2024. Gemma. https://doi.org/10.34740/KAGGLE/M/3301
[32] Jaromír Šavelka and Kevin D. Ashley. 2015. Transfer of predictive models for classification of statutory texts in multi-jurisdictional settings. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (San Diego, California) *(ICAIL '15)*. Association for Computing Machinery, New York, NY, USA, 216–220. https://doi.org/10.1145/2746090.2746109
[33] Bernhard Waltl, Johannes Muhr, Ingo Glaser, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2017. Classifying legal norms with active machine learning. In *Legal Knowledge and Information Systems.* IOS Press, 11–20.
[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6