Fine-tuning GPT-3 for legal rule classification

Davide Liga[1] and Livio Robaldo[2]

[1]University of Luxembourg (davide.liga@uni.lu)
[2]Swansea University (livio.robaldo@swansea.ac.uk)

Abstract

In this paper, we propose a Legal Rule Classification (LRC) task using one of the most discussed language model in the field of Artificial Intelligence, namely GPT-3, a generative pretrained language model. We train and test the proposed LRC task on the GDPR encoded in LegalDocML (Palmirani and Vitali, 2011) and LegalRuleML (Athan et al., 2013), two widely used XML standards for the legal domain. We use the LegalDocML and LegalRuleML annotations provided in (Robaldo, Bartolini, and Lenzini, 2020) to fine-tuned GPT-3. While showing the ability of large language models (LLMs) to easily learn to classify legal and deontic rules even on small amount of data, we show that GPT-3 can significantly outperform previous experiments on the same task. Our work focused on a multiclass task, showing that GPT-3 is capable to recognize the difference between obligation rules, permission rules and constitutive rules with performances that overcome previous scores in LRC.

1 Introduction

Recent developments in Artificial Intelligence (AI), in the field of Natural Language Processing (NLP) as well as in other fields such as Computer Vision and Speech Recognition, have shown the groundbreaking power of pre-trained Large Language Models (LLMs).

The success of LLMs started in nearly 2018, with Google's BERT (Devlin et al., 2018) and OpenAI's GPT (Radford and Narasimhan, 2018). The idea behind LLMs is straightforward: we can create powerful language models by training neural architectures on huge amounts of data. In this regard, the neural architecture from which many LLMs derives is the transformer architecture (Vaswani et al., 2017). These language models have shown the ability to be applied to a wide range of NLP tasks, overcoming the state of the art in many NLP challenges.

In this paper, we focus on Legal Rule Classification (LRC), which has been tackled only by few scholars in the field of Artificial Intelligence and Law (AI&Law), despite its importance (Robaldo et al., 2019).

We compare the performance of the few previous studies that employ LLMs with the performance of our work, which employs two fine-tuned version of GPT-3. To the best of our knowledge, this paper is the first attempt to fine-tune the well-known OpenAI's

GPT-3 (currently, the most powerful language model in the world) on legal data, and in particular on the recognition of deontic rules.

The capability of automatically detecting deontic rules from natural language is an important research direction for the whole AI&Law community. It will enable the development of advanced legal expert systems (cf. (Boella et al., 2016)); in addition, the use LLMs will specifically foster a deeper integration of "bottom-up" data-driven AI with "top-down" symbolic AI, i.e., an integration of LLMs with the most recent theoretical results in formal deontic logic and argumentation (Wyner and Peters, 2011; Ashley, 2017; Sun and Robaldo, 2017). In this work, we show:

- How GPT-3 can achieve remarkable results in recognising legal and deontic rules
- How existing symbolic knowledge can be employed to further exploit GPT-3 potential

Although the idea of using AI to automatically extract rules and deontic modalities from legal text is not new (see Section 2), there are some important obstacles which usually prevented the AI&Law community from achieving better results, namely the lack of available data designed *ad hoc* for the classification of rules and deontic modalities. In fact, annotating and creating this kind of data is not just time-consuming but also costly, and it requires domain experts, who are not always available.

On the other hand, the creation of high-quality datasets is usually in trade-off with the size of these datasets, which in turn limit the accuracy of standard Machine Learning classifiers (e.g., (Boella et al., 2013)), especially those employing deep neural architectures (e.g., (Song et al., 2022)), which notoriously need huge amounts of data.

In this regard, LLMs recently paved the way towards a new paradigm in AI, sometimes referred to as "Transfer Learning", which indicates the idea that we can use LLMs by transferring what they "learnt" during their pre-training phase to downstream tasks and downstream data. This showed the ability of these models to achieve impressive results even on small datasets. For these reasons, many researchers started using LLMs like BERT (Devlin et al., 2018) which is one of the most famous examples of successful pre-trained neural architectures, used in many downstream tasks (including tasks which employed very small datasets (Liga and Palmirani, 2020)).

In the past years, some legal XML standards have been proposed, which are capable of providing researchers with machine-readable legal knowledge. The most popular one is undoubtedly Akoma Ntoso, a.k.a. LegalDocML[1], which is used to represent legal documents in XML format, thus allowing to encode the structure of legal documents (sections, preambles, articles, etc.) as well as metadata related to the nature and the history of such documents. Another legal XML standard is LegalRuleML[2], which is capable of representing the logical dimension of legal documents, including logical deontic rules, related for example to obligations and permissions.

In this work, we want to use these XML standards in combination with one of the most PML in the world, and probably the most popular so far, namely GPT-3. Specifically, this paper will show the potential of using legal XML documents as source of data for applying GPT-3 on downstream tasks such as LRC.

[1] https://www.oasis-open.org/committees/legaldocml

[2] https://www.oasis-open.org/committees/legalruleml

As said above, LRC, and in particular the automatic classification of deontic rules from natural language, is a task which has received little attention by the AI&Law community, despite its usefulness within legal expert systems. This task consists in classifying single legal sentences or single legal provisions as containing deontic modalities such as Obligations, Prohibitions and Permissions (Nguyen et al., 2022) (Liga and Palmirani, 2022b).

In the Section 2, we will describe some related works, while Section 3 will shortly describe the role of LegalXML standards. In Section 4, we will give a brief overview of our method, including a short introduction on both the data extraction technique (2.1) and the classification technique (2.2). In the following two sections, we will give a more exhaustive description of the retrieved data (Section 5), and a detailed report on the experimental settings with their respective results (Section 6). In particular, we present four experimental scenarios and two prompt strategies, i.e., eight settings in total. Finally, Section 7 concludes the paper.

2 Related Works

There have been (relatively few) attempts in literature to employ NLP methodologies to automatically detect rules in legal documents. Among these first attempts, one can find studies tackling the classification of deontic elements (Dragoni et al., 2018) as parts of a wider range of targets (Kiyavitskaya et al., 2008) (Waltl et al., 2017), (Gao and Singh, 2014), (de Maat and Winkels, 2008). Among these first attempts to classify obligations from legal texts there is (Kiyavitskaya et al., 2008), which focused on the regulations of Italy and US. Their method employed word lists, grammars and heuristics to extract obligations among other targets such as rights and constraints. Another work which tackled the classification of deontic statements is (Waltl et al., 2017), which focused on the German tenancy law and classified 22 classes of statements (among which there were also prohibitions and permissions). The method used active learning with multinomial naive bayes, logistic regression and multi-layer perceptron classifiers, on a corpus of 504 sentences. In (Gao and Singh, 2014), the authors used machine learning to extract six classes of normative relationships: prohibitions, authorizations, sanctions, commitments and powers.

As mentioned above, the first studies which tackled the classification of deontic rules focused on deontic elements as parts of a wider range of targets. Perhaps the first study which mainly focused on the deontic sphere is (Neill et al., 2017). This work was focused on the financial legislation to classify legal sentences using a Bi-LSTM architecture, with a training dataset containing 1,297 instances (596 obligations, 94 prohibitions, and 607 permissions).

While the majority of the studies related to LRC are strongly focused symbolic and rule-based Artificial Intelligence (Wyner and Peters, 2011), (de Maat and Winkels, 2008), some studies instead focused on the use of the neural networks (Neill et al., 2017), (Chalkidis, Androutsopoulos, and Michos, 2018) and LLM (Shaghaghian et al., 2020), (Joshi, Anish, and Ghaisas, 2021), (Liga and Palmirani, 2022b). To the best of our knowledge, the first study which employed a language model (i.e., BERT) for the classification of deontic sentences are (Shaghaghian et al., 2020) and (Joshi, Anish, and

Ghaisas, 2021). (Shaghaghian et al., 2020) used four pre-trained architectures (BERT, DistilBERT, RoBERTa, and ALBERT) but focused just on the binary detection duties vs non-duties. (Joshi, Anish, and Ghaisas, 2021) also focused on permissions, achieving an average precision and recall of 90% and 89.66% respectively. A recent work, from which this paper is inspired, showed how to use BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019) and LegalBERT (Chalkidis et al., 2020) to classify obligations, permissions and constitutive rules automatically (Liga and Palmirani, 2022b), inspired in turn by previous positive results using Tree Kernel algorithms (Liga and Palmirani, 2022a)

Among these studies, no one attempted to use generative LLM, which recently gained immense success thanks to the release of GPT-3 and ChatGPT (Brown et al., 2020). In this work, we want to cover this gap, fostered also by the fact that GPT-3 is currently the most powerful language model, or at least the one which received most enthusiastic attention by media and scientific communities.

Therefore, similarly to (Liga and Palmirani, 2022b), our work presents a transfer learning approach of machine learning, which combines the symbolic information of legal XML formats with the sub-symbolic power provided by GPT-3. By leveraging the information channeled by the biggest LegalRuleML knowledge base available, we present four different scenarios of classification:

1. Rule vs Non-rule
2. Deontic vs Non-deontic
3. Obligation vs Permission vs Non-deontic
4. Obligation vs Permission vs Constitutive Rule vs Non-rule

The novelty and the power of generative AI methodologies jointly with the combined use of LegalDocML and LegalRuleML are two major contributions of this study, along with the design of the experimental settings in four different classification scenarios using two different prompt strategies. In addition, legal XML formats such as LegalDocML and LegalRuleML are usually written and validated by legal experts. In other words, for the task of detecting deontic classes, the extraction of data from this kind of documents can arguably offer a more convenient and robust solution compared to the use of general-purpose datasets which might be only partially related to the deontic sphere.

3 LegalXML

This work exploits the potential of two renowned LegalXML standards: LegalDocML and LegalRuleML, which are key in the field of AI&Law and Legal Knowledge Representation.

LegalDocML applies the XML standard specifically to legal documents. This system encodes legal and legislative documents into a structured, machine-readable format, thereby facilitating their management and exchange across various systems and platforms. LegalDocML is highly valued for its ability to streamline the processing and analysis of legal documents. This is achieved by standardizing the format of these documents, allowing them to be interpreted, displayed, and stored uniformly across any system. Furthermore, LegalDocML bolsters the accuracy and efficiency of legal document comparisons and improves their interoperability. Consequently, this facilitates legal practitioners, institutions, and software developers to share and utilize information harmoniously in a standardized format.

LegalRuleML encompasses features for representing the norms and rules in legal documents. Much like LegalDocML, LegalRuleML is also XML-based, allowing the embedding of rules and logic into a machine-readable format. LegalRuleML is advantageous as it enables automated reasoning about the content of legal documents from a normative provisions perspective. It assists in managing and manipulating large volumes of legal norms, linking rules to relevant legal or legislative documents based on underlying legal theories. Essentially, LegalRuleML renders the normative dimension of legal documents computationally understandable, thus enhancing accessibility and facilitating the development of advanced legal tech solutions. These include document automation, knowledge extraction, and legal texts analysis. Through this, systems can ensure a consistent interpretation and application of laws, legal rulings, and policies, thereby boosting the efficiency and accuracy of automated legal systems.

4 Methodology

The objective of this work is twofold. On the one hand, it aims at showing that LegalDocML and LegalRuleML can be combined to feed generative AI machine learning algorithms with reliable data for the classification of deontic rules. On the other hand, it aims at testing the use of transfer learning on the task of deontic rule classification. The first aspect (i.e., the combination of LegalDocML and LegalRuleML) is related to the methodology that has been used to extract the legal knowledge and data. The second aspect (i.e., the usage of transfer learning as machine learning algorithm) is related to the methodology for the classification. The combination of these two methodological aspects (i.e., method of data/knowledge extraction and method of classification) can be defined as a Hybrid AI approach, since it combines symbolic knowledge with sub-symbolic knowledge (Gomez-Perez, Denaux, and Garcia-Silva, 2020), (Rodríguez-Doncel et al., 2020), (Liga, 2022).

4.1 Data Extraction Method

Regarding the methodology, we replicated the approach proposed in (Liga and Palmirani, 2022a) and (Liga and Palmirani, 2022b), where some machine learning classifiers were trained, in supervised settings, on the information contained in two legal XML formats, namely LegalDocML and LegalRuleML.

As in (Liga and Palmirani, 2022b), a major contribution of this work is on the novelty and the power of transfer learning methodologies jointly with the combined use of LegalDocML and LegalRuleML[3].

Similarly to (Liga and Palmirani, 2022a) and (Liga and Palmirani, 2022b), this work used a dataset of 707 atomic legal provisions found in the European General Data Pro-

[3] Also, legal XML formats such as LegalDocML and LegalRuleML are documents which are written by legal experts, providing machine learning algorithms with high quality data.

tection Regulation (GDPR). The data was obtained from the DAPRECO Knowledge Base (D-KB for short) (Robaldo et al., 2020), which is the largest knowledge base in LegalRuleML as well as the largest knowledge base formalized in reified Input/Output Logic (Robaldo, 2021).

The D-KB represents the logical-deontic dimension of the GDPR and it currently contains 966 reified I/O logic formulas, including 271 obligations, 76 permissions, and 619 constitutive rules. Constitutive rules are used to trigger specific inferences for the modeled rules and are distinct from obligations or permissions in that they do not convey information about deontic modalities. DAPRECO also includes connections between each formula and its corresponding structural portion of the legal document in the LegalDocML representation of the GDPR, allowing for a link between the logical-deontic aspects of legal documents, represented by the 966 Input/Output formulas in DAPRECO, and the natural language statements in the legal text, represented by the LegalDocML representation of the GDPR. The combination of LegalDocML and LegalRuleML is also helpful in reconstructing the exact natural language target, such as by combining lists of obligations in legal texts into a single sentence. In this regard, the role of LegalDocML is particularly important, allowing for the reconstruction of atomic provisions (e.g., when a provision is split into a list).

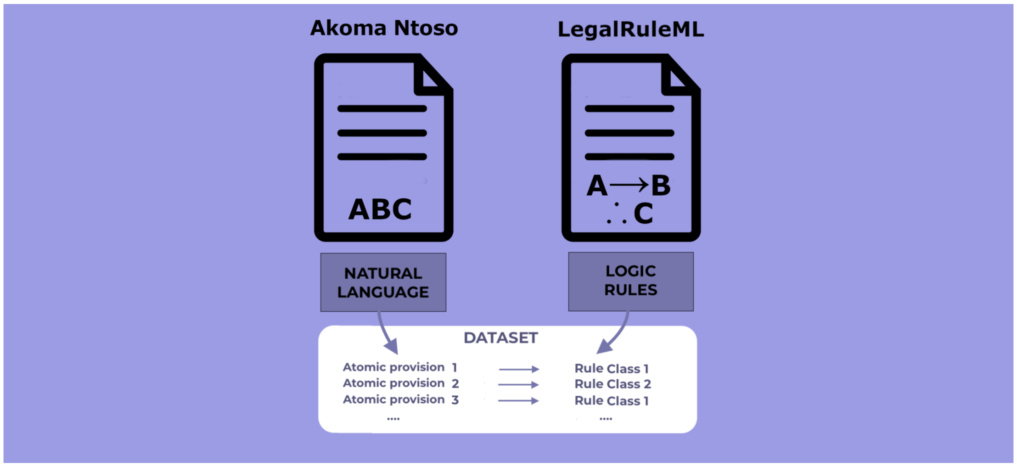


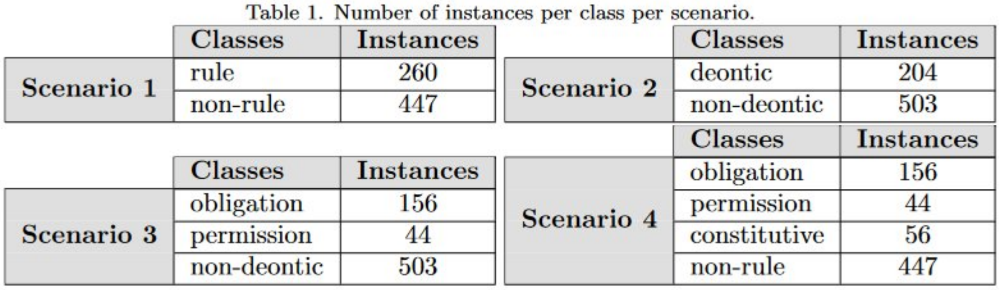
Figure 1: Knowledge extraction from LegalDocML and LegalRuleML. Note that each extracted instance refers to an atomic normative provision (generally contained in paragraphs or points), and may sometimes consist of more than one sentence.

By combining the structural information from LegalDocML and the deontic information in LegalRuleML we extracted 707 labelled legal provisions in total. The labels of these sentences are the same as those provided by DAPRECO with the addition of a "non-rule" category. We abbreviated "obligationRule", "permissionRule", "constitutiveRule" in "obligation", "permission" and "constitutive" respectively.

The class "obligation" is referred to those sentences which have at least one obligation rule in their related formulæ. The class "permission" is referred to those sentences which have at least one permission rule in their related formulæ. The class “constitutive” is referred to those sentences which just constitutive rules in their related formulæ. We also considered a class “non-rule” which is referred to all sentences which have no legal rule at all, and “non-deontic” which is referred to all sentences which does not contain neither obligations nor permissions (they may still contain constitutive rules, though)[4].

These labels allow for four different experimental settings, as shown in Table 1, which provide different levels of granularity as shown in Figure 2.

Table 1. Number of instances per class per scenario.



	Classes	Instances
Scenario 1	rule	260
	non-rule	447

	Classes	Instances
Scenario 2	deontic	204
	non-deontic	503

	Classes	Instances
Scenario 3	obligation	156
	permission	44
	non-deontic	503

	Classes	Instances
Scenario 4	obligation	156
	permission	44
	constitutive	56
	non-rule	447

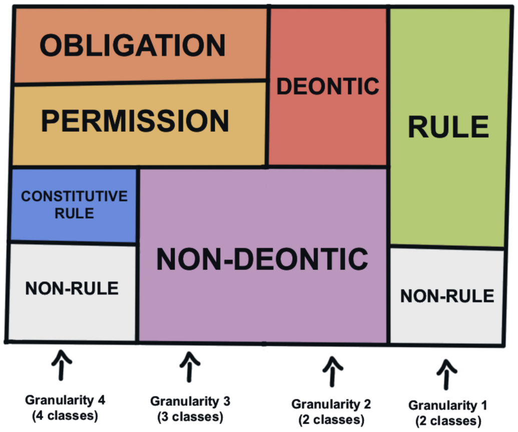

OBLIGATION
PERMISSION
DEONTIC
RULE
CONSTITUTIVE RULE
NON-RULE
NON-DEONTIC
NON-RULE
Granularity 4 (4 classes)
Granularity 3 (3 classes)
Granularity 2 (2 classes)
Granularity 1 (2 classes)

Figure 2: Levels of granularity of the four classification scenarios.

4.2 Classification method

Regarding the transfer learning classification method, we fine-tuned GPT-3 (Brown et al., 2020) on our dataset of 707 extracted legal provisions. For the fine-tuning approach we engineered two different simple prompts to give GPT-3 the necessary instructions for classifying the atomic provisions’ classes.

Our first prompt has the following template:

[4] For the multi-classifications (i.e. Scenario 3 and 4) four statements have been removed, since the classes “obligation” and “permission” overlapped.

prompt: “[ATOMIC LEGAL PROVISION] ->”

completion: “ [CLASS AS NUMBER]”

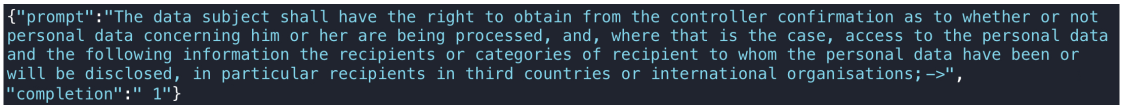
Where “[ATOMIC LEGAL PROVISION]” is the single atomic provision extracted from LegalDocML and LegalRuleML, while“->” is our classification marker, while “\n” stands for a new line. The completion of this prompt is the class represented as a number (in the case of scenario 4, numbers 0, 1, 2, 3 stand for “none”, “obligation”, “permission” and “constitutive” respectively).

We also tested a second prompt, namely:

prompt: “[ATOMIC LEGAL PROVISION]\n\nThe previous text is a ->”

completion: “ [CLASS NAME]”

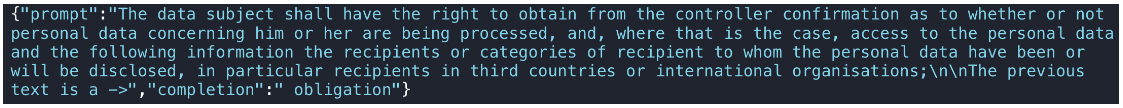
This time, the completion of this prompt is the class of the atomic legal provision represented as words (not as numbers). An example of legal provision marked using prompt 1 is the following:



```
{"prompt":"The data subject shall have the right to obtain from the controller confirmation as to whether or not
personal data concerning him or her are being processed, and, where that is the case, access to the personal data
and the following information the recipients or categories of recipient to whom the personal data have been or
will be disclosed, in particular recipients in third countries or international organisations;->",
"completion":" 1"}
```

Figure 3: Example of legal provision with prompt 1

An example of legal provision marked using prompt 2 is the following:



```
{"prompt":"The data subject shall have the right to obtain from the controller confirmation as to whether or not
personal data concerning him or her are being processed, and, where that is the case, access to the personal data
and the following information the recipients or categories of recipient to whom the personal data have been or
will be disclosed, in particular recipients in third countries or international organisations;\n\nThe previous
text is a ->","completion":" obligation"}
```

Figure 4: Example of legal provision with prompt 2

5 Data

In this study, we employed a dataset containing a total amount of 707 atomic normative provisions[5] extracted from the European GDPR (General Data Protection Regulation). In particular, as already explained above, we gathered the normative provisions of the GDPR by employing the D-KB (Robaldo et al., 2020), which represents in LegalRuleML the norms of the GDPR. The current version of the D-KB[6] includes 966 formulas in reified Input/Output logic: 271 obligations, 76 permissions, and 619 constitutive rules. As explained in (Robaldo et al., 2020), the number of constitutive rules is much higher

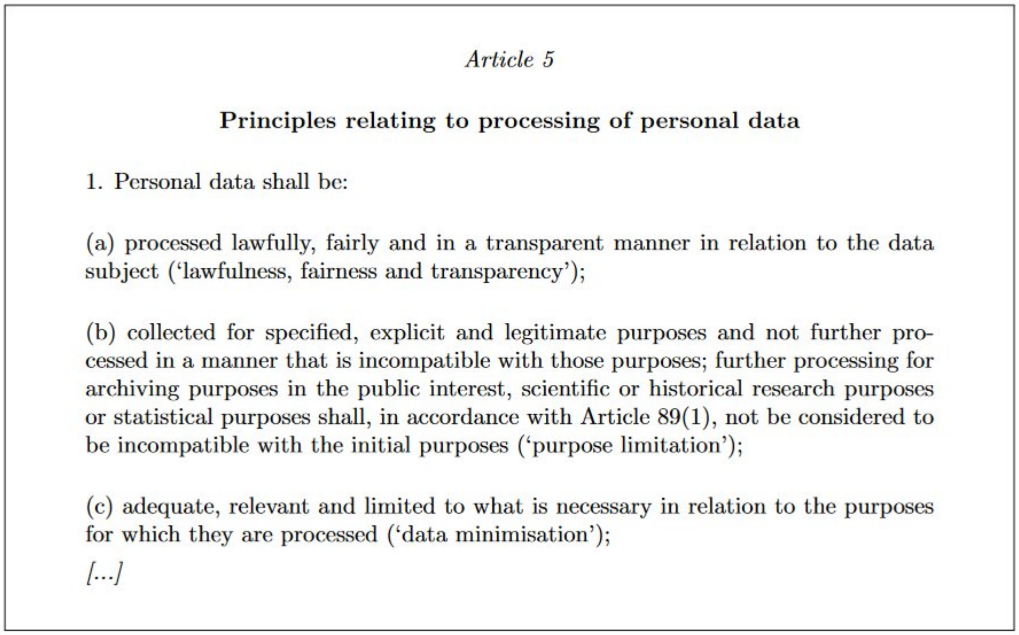
[5] For the selection of the provisions, we excluded preamble and conclusion from the main legal document of the GDPR, thus keeping just the provisions within the body of the GDPR. These provisions are generally paragraphs or list points, and may sometimes consist of more than one sentence.

[6] The D-KB can be freely downloaded from: `https://github.com/dapreco/daprecokb/blob/master/gdpr/rioKB_GDPR.xml`

than permissions and obligations because constitutive rules are needed to trigger special inferences for the modelled rules. In other words, while constitutive rules are an indicator of the existence of a rule, they do not give information about deontic modalities.

An important aspect of the D-KB is that it contains the references to the structural elements (paragraphs, point, etc.) where each norm of the GDPR, represented in the D-KB as an if-then rule, can be found. More precisely, it refers to the LegalDocML representation of the GDPR, where all structural elements of the GDPR can be found following the LegalDocML standards[7]. In other words, using a LegalRuleML knowledge base like the D-KB and the corresponding LegalDocML representation, it is possible to connect the logical-deontic sphere of legal documents to the natural language statements in the legal text, provided by the LegalDocML representation of the GDPR.

It is important to remark that this combination of LegalDocML and LegalRuleML also facilitates the reconstruction of the exact target, in terms of natural language, where each provision is located. For example, many obligations of legal texts are split into lists, and LegalDocML is crucial to reconstruct those pieces of natural language into a unique piece of natural language. For example, Article 5 of the GDPR[8] states:



Article 5

Principles relating to processing of personal data

1. Personal data shall be:

(a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

[...]

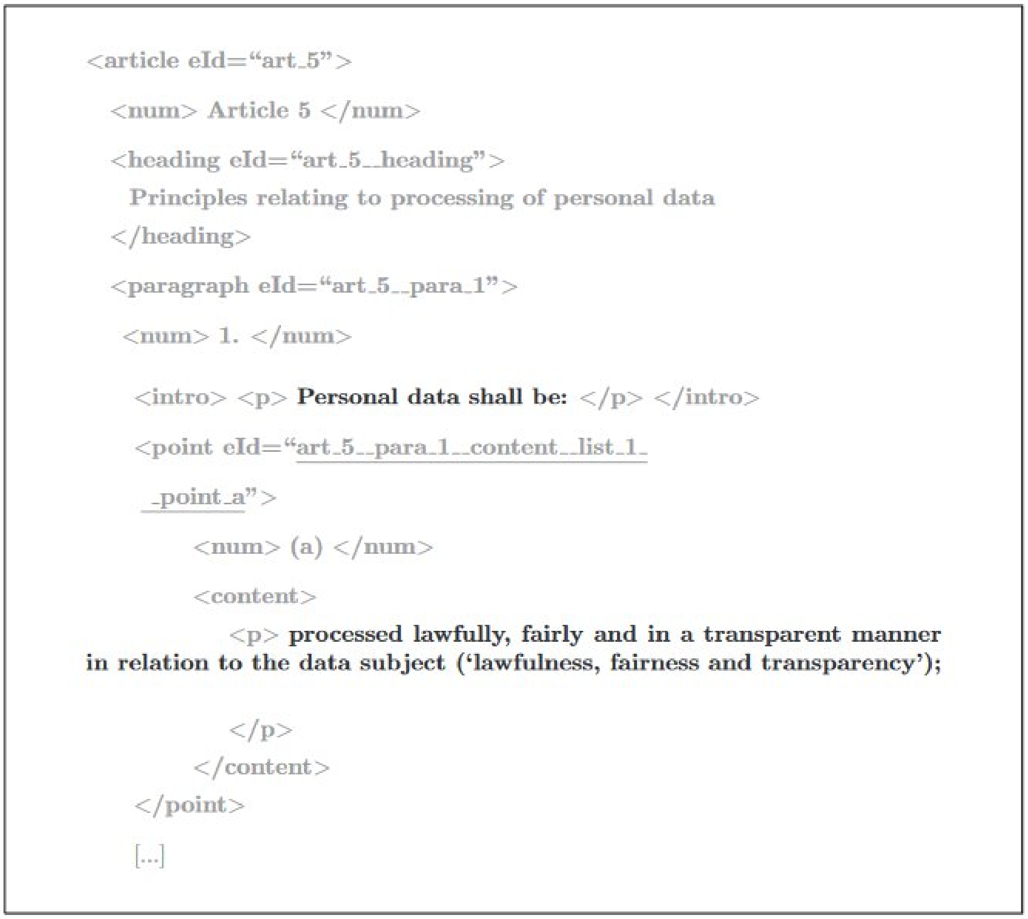
As can be seen in the text above, *paragraph 1* of *Article 5* is a list composed of an introductory part ("Personal data shall be:") and different points. For simplicity, we only reported the first three points of *paragraph 1*, namely *point a*, *point b* and *point c*. From the point of view of the natural language, each deontic sentence is split between

[7] The LegalDocML representation of the GDPR can be currently accessed and downloaded from `https://github.com/guerret/lu.uni.dapreco.parser/blob/master/resources/akn-act-gdpr-full.xml`.

[8] `https://eur-lex.europa.eu/eli/reg/2016/679/oj\#d1e1807-1-1`.

the introductory part (which contains the main deontic verb "shall") and the text of each point. While the introductory part contains the main deontic verb, the actual deontic information is contained within each point.

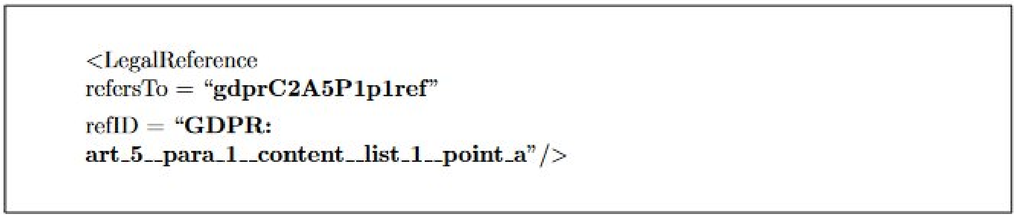
The LegalDocML formalization for *point a* is the following:



```
<article eId="art_5">
  <num> Article 5 </num>
  <heading eId="art_5__heading">
    Principles relating to processing of personal data
  </heading>
  <paragraph eId="art_5__para_1">
   <num> 1. </num>
    <intro> <p> Personal data shall be: </p> </intro>
    <point eId="art_5__para_1__content__list_1_
     _point_a">
        <num> (a) </num>
        <content>
            <p> processed lawfully, fairly and in a transparent manner
in relation to the data subject ('lawfulness, fairness and transparency');
            </p>
        </content>
    </point>
    [...]
```

In the D-KB, which employs a LegalRuleML, a series of <Legal Reference> elements can be found, which contain the structural portion where the deontic formulas are located, referenced by using the LegalDocML naming convention[9]. For example, the reference of the above mentioned *point a* is encoded in the D-KB as follows:

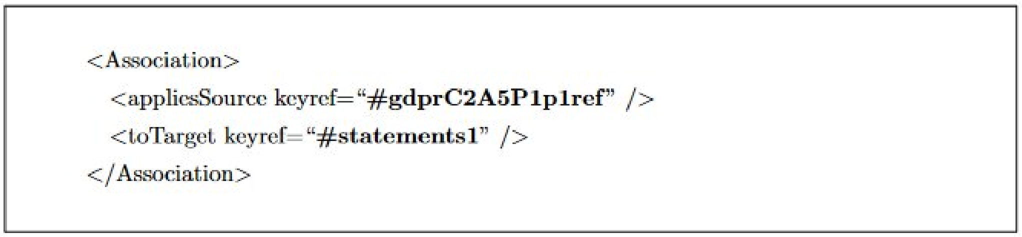
[9] https://docs.oasis-open.org/legaldocml/akn-nc/v1.0/csprd01/akn-nc-v1.0-csprd01.html.



```
<LegalReference
refersTo = “gdprC2A5P1p1ref”
refID = “GDPR:
art_5__para_1__content__list_1__point_a”/>
```

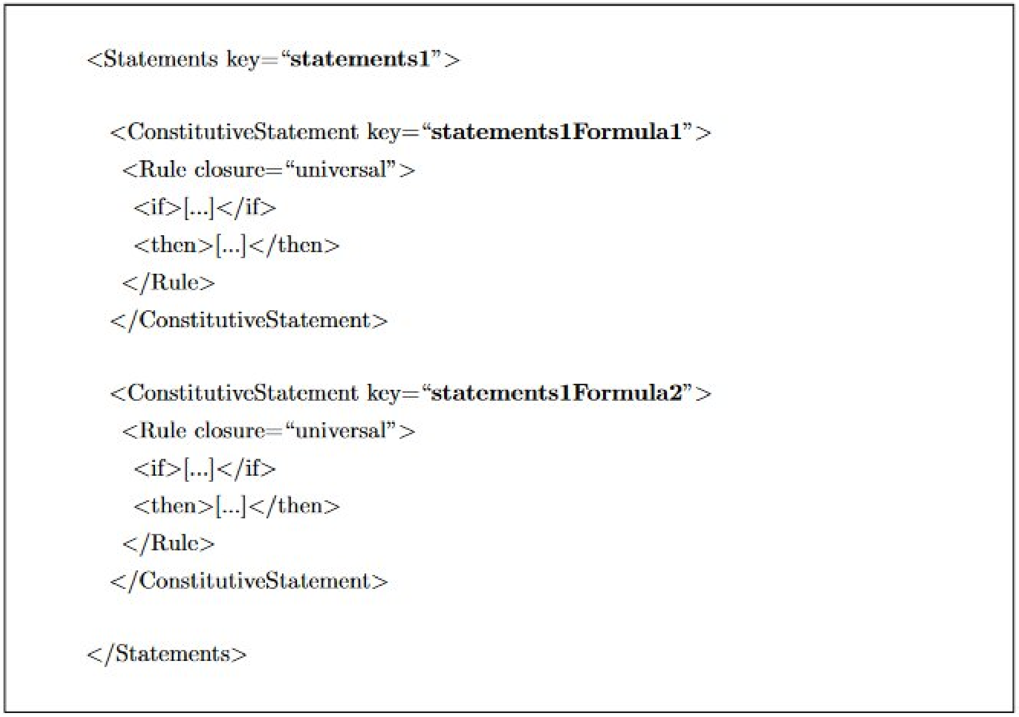
in which the “refersTo” attribute indicates the internal ID of the reference, and the “refID” attribute indicates the external ID of the reference using the LegalDocML naming convention. The prefix “GDPR” stands for the LegalDocML uri of the GDPR, namely “/akn/eu/act/regulation/2018-05-25/eng@2018-05-25/!main#”.

In turn, this <LegalReference> element is then associated to its target group of logical statements, which collects the group of logical formulas related to this legal reference (so, in this case, related to *point a* of the first paragraph of *Article 5*). Such association is modelled as follows:



```
<Association>
  <appliesSource keyref=“#gdprC2A5P1p1ref” />
  <toTarget keyref=“#statements1” />
</Association>
```

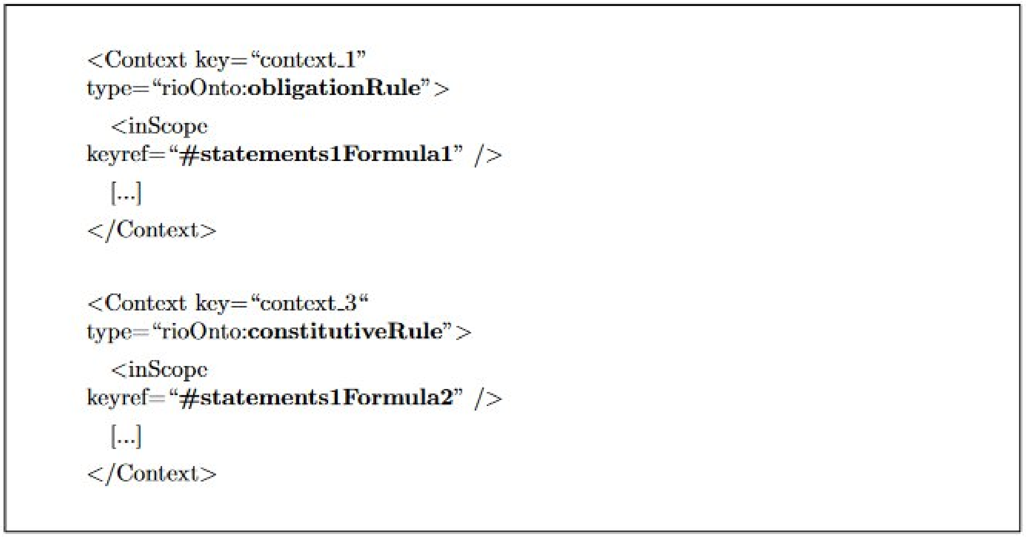
Where the attribute “keyref” of the target connects the source to the collection of statements whose “key” attribute is “statements1”:



```
<Statements key="statements1">

  <ConstitutiveStatement key="statements1Formula1">
   <Rule closure="universal">
     <if>[...]</if>
     <then>[...]</then>
   </Rule>
  </ConstitutiveStatement>

  <ConstitutiveStatement key="statements1Formula2">
   <Rule closure="universal">
     <if>[...]</if>
     <then>[...]</then>
   </Rule>
  </ConstitutiveStatement>

</Statements>
```

It is important to underline that each natural language statement can have multiple formulas in the logical sphere. For this reason, the element <Statements> here shows a collection of two logical formulas.

To finally associate the portion of natural language extracted from LegalDocML to a class related to the logical sphere (i.e. the deontic class), one must look at the <Context> elements which are related to the two formulas we found.



```
<Context key=“context_1”
type=“rioOnto:obligationRule”>
  <inScope
keyref=“#statements1Formula1” />
  [...]
</Context>

<Context key=“context_3“
type=“rioOnto:constitutiveRule”>
  <inScope
keyref=“#statements1Formula2” />
  [...]
</Context>
```

As can be seen from the text above, the first formula (which is called here “statements1Formula1”) is associated with the ontological class “obligationRule”, while the second formula (which is called “statements1Formula2”) is associated with the ontological class “constitutiveRule”. This means that the piece of natural language expressed in *point a* of the first paragraph of *Art. 5* of the GDPR contains, at the logical level, a constitutive rule and an obligation rule.

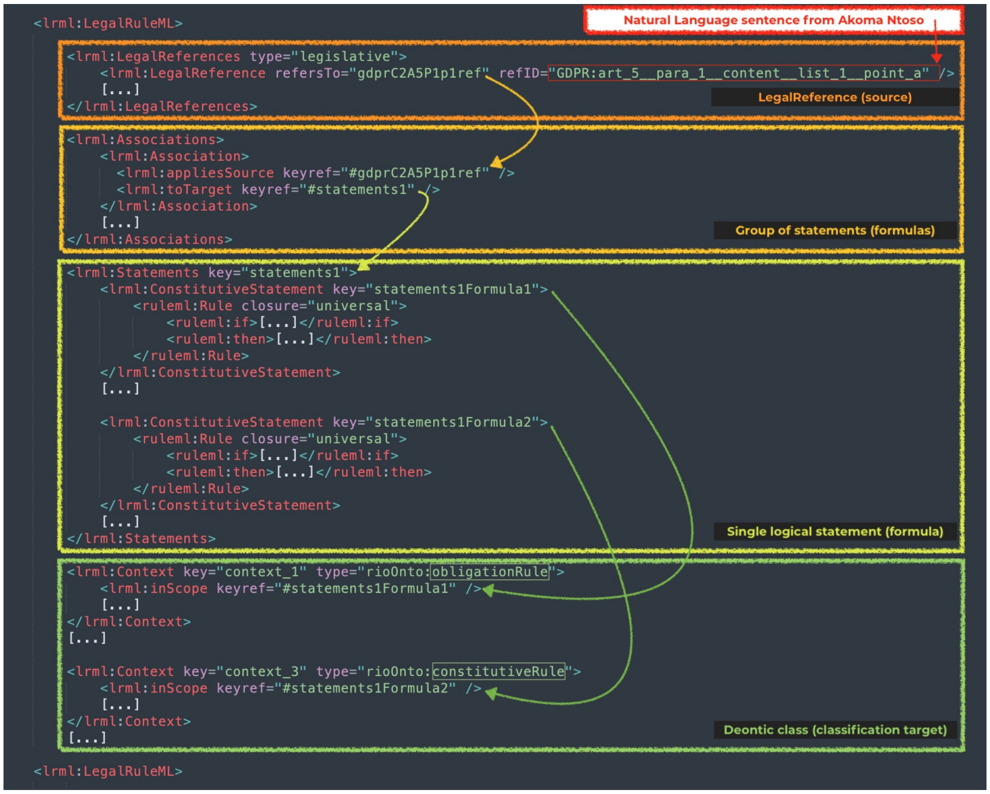
Figure 6 shows the full series of steps from the natural language sphere (located in the LegalDocML) to the logical sphere (i.e. the LegalRuleML formalization) where the deontic classes are located. The figure explains step-by-step how the combination of LegalDocML and LegalRuleML helped us in the extraction of annotated labelled data.

6 Experiment settings and results

As far as the experimental settings are concerned, the dataset was divided into training and validation sets, with a 80/20 split[10]. Moreover, we employed Ada as engine for GPT-3 we employed Ada, the standard choice for classification tasks. Also, we noticed empirically that GPT-3's Ada outperforms Davinci in simple classification tasks, while Davinci is more appropriate in generative tasks. We reported the results of BERT (Devlin et al., 2018), distilBERT (Sanh et al., 2019) and LegalBERT (Chalkidis et al., 2020), taken from (Liga and Palmirani, 2022b), which we used as baselines, and we also reported our two fine-tuned GPT-3 models. These two fine-tuned language models has been achieved after 4 epochs of fine-tuning. As far as the hyper-parameters are concerned, we set the learning rate multiplier at 0.1, the prompt loss weight at 0.01 and the batch size at 1. The final results on the validation set are reported in Table 2 and 3, where it can be seen that all fine-tuned GPT-3 models outperforms the previous results taken from (Liga

[10] Note that this ratio was slightly different in (Liga and Palmirani, 2022b).

and Palmirani, 2022b), both in terms of F1 scores (Table 2) and in terms of accuracy (Table 3). Figure 6 shows a graph with accuracy and F1 scores for our two fine-tuned models.



```
<lrml:LegalRuleML>
    <lrml:LegalReferences type="legislative">
        <lrml:LegalReference refersTo="gdprC2A5P1p1ref" refID="GDPR:art_5__para_1__content__list_1__point_a" />
        [...]
    </lrml:LegalReferences>
    <lrml:Associations>
        <lrml:Association>
          <lrml:appliesSource keyref="#gdprC2A5P1p1ref" />
          <lrml:toTarget keyref="#statements1" />
        </lrml:Association>
        [...]
    </lrml:Associations>
    <lrml:Statements key="statements1">
        <lrml:ConstitutiveStatement key="statements1Formula1">
            <ruleml:Rule closure="universal">
                <ruleml:if>[...]</ruleml:if>
                <ruleml:then>[...]</ruleml:then>
            </ruleml:Rule>
        </lrml:ConstitutiveStatement>
        [...]

        <lrml:ConstitutiveStatement key="statements1Formula2">
            <ruleml:Rule closure="universal">
                <ruleml:if>[...]</ruleml:if>
                <ruleml:then>[...]</ruleml:then>
            </ruleml:Rule>
        </lrml:ConstitutiveStatement>
        [...]
    </lrml:Statements>
    <lrml:Context key="context_1" type="rioOnto:obligationRule">
        <lrml:inScope keyref="#statements1Formula1" />
        [...]
    </lrml:Context>
    [...]

    <lrml:Context key="context_3" type="rioOnto:constitutiveRule">
        <lrml:inScope keyref="#statements1Formula2" />
        [...]
    </lrml:Context>
    [...]

<lrml:LegalRuleML>
```

Natural Language sentence from Akoma Ntoso
LegalReference (source)
Group of statements (formulas)
Single logical statement (formula)
Deontic class (classification target)

Figure 5: Description of the process of class extraction from LegalDocML and DAPRECO.

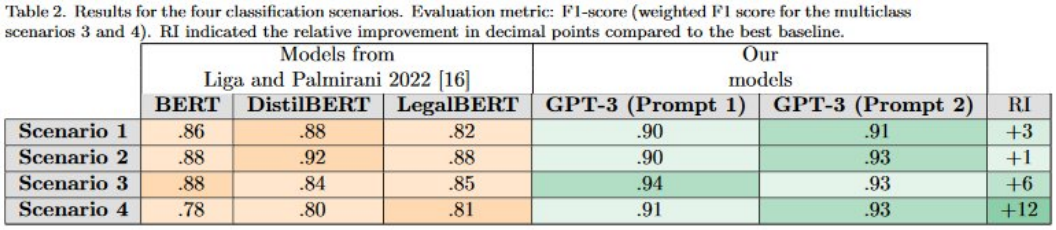
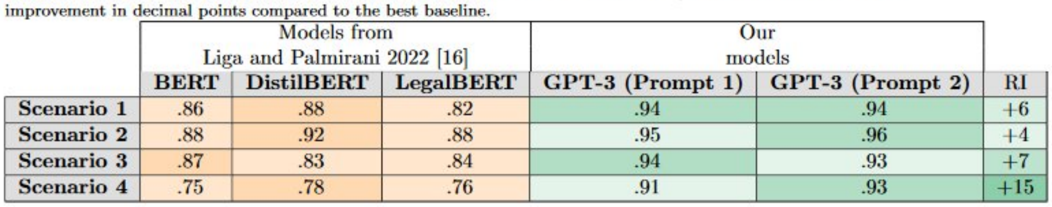


Table 2. Results for the four classification scenarios. Evaluation metric: F1-score (weighted F1 score for the multiclass scenarios 3 and 4). RI indicated the relative improvement in decimal points compared to the best baseline.

	Models from Liga and Palmirani 2022 [16]			Our models		
	BERT	**DistilBERT**	**LegalBERT**	**GPT-3 (Prompt 1)**	**GPT-3 (Prompt 2)**	RI
Scenario 1	.86	.88	.82	.90	.91	+3
Scenario 2	.88	.92	.88	.90	.93	+1
Scenario 3	.88	.84	.85	.94	.93	+6
Scenario 4	.78	.80	.81	.91	.93	+12

Table 3. Results for the four classification scenarios. Evaluation metric: Accuracy. RI indicated the relative improvement in decimal points compared to the best baseline.



	Models from Liga and Palmirani 2022 [16]			Our models		
	BERT	**DistilBERT**	**LegalBERT**	**GPT-3 (Prompt 1)**	**GPT-3 (Prompt 2)**	RI
Scenario 1	.86	.88	.82	.94	.94	+6
Scenario 2	.88	.92	.88	.95	.96	+4
Scenario 3	.87	.83	.84	.94	.93	+7
Scenario 4	.75	.78	.76	.91	.93	+15

Regarding scenario 1 (i.e., the binary classification between rules and non-rules) we achieved .91 of F1 score with prompt 2 and .90 with prompt 1, while the best previous result was .88 with DistilBERT.

In the second scenario, which consists in performing a binary classification between deontic rules (obligations and permissions) vs non-deontic rules (i.e., everything which is neither an obligation nor a permission), GPT-3 reached .93 of F1 score, against .92 (DistilBERT). This is the only scenario where the baseline competed with GPT-3.

Regarding the third scenario, consisting in a multi-class classification involving obligations and permissions against "non-deontic" (i.e., everything else which is neither an obligation nor a permission), GPT-3 reached .94 of F1 score, against .88 (BERT).

The most significant improvement is related to the most difficult scenario, the fourth one, which involves the most granular classification, considering all the available classes (obligation, permission, constitutive rule, non-rule). While LegalBERT, the previous highest result, achieved .81 of weighted F1 score, our fine-tuned GPT-3 models reached .93 of weighted F1 score, more than one decimal point.

As can be seen from Table 3, the relative improvement from the baseline is more evident when considering accuracy as evaluation metric.

7 Conclusion and Future works

In this work, we showed that autoregressive generative LLMs like GPT-3 can outperform auto-encoding LLMs like BERT achieving higher performances in the task of Legal Rule Classification (LRC). We performed LRC on four multiclass classification scenarios, involving obligations, permissions, constitutive rules and "non-rules" (i.e., legal provisions which do not contain any kind of rule).

Importantly, our work shows how Hybrid AI approaches can successfully combine symbolic and sub-symbolic Artificial Intelligence. On the one side, we provide legal expert knowledge encoded in LegalXML formats such as LegalRuleML and Akoma Ntoso. On the other side, we leverage the power of GPT-3, arguably the most powerful language model currently available for fine-tuning in the field of Artificial Intelligence. This Hybrid AI approach shows the potential of combining top-down (knowledge-driven) approaches with bottom-up (data-driven) methods.

To the best of our knowledge, this paper is the first to fine-tune GPT-3 (currently one of the most powerful and intensively discussed models) on legal data. Although our work showcases the impressive capabilities of LLMs to learn from small quantities of data, as well as the superiority of GPT-3 over other LLMs, more experiments are required to validate these trends using different legal datasets.

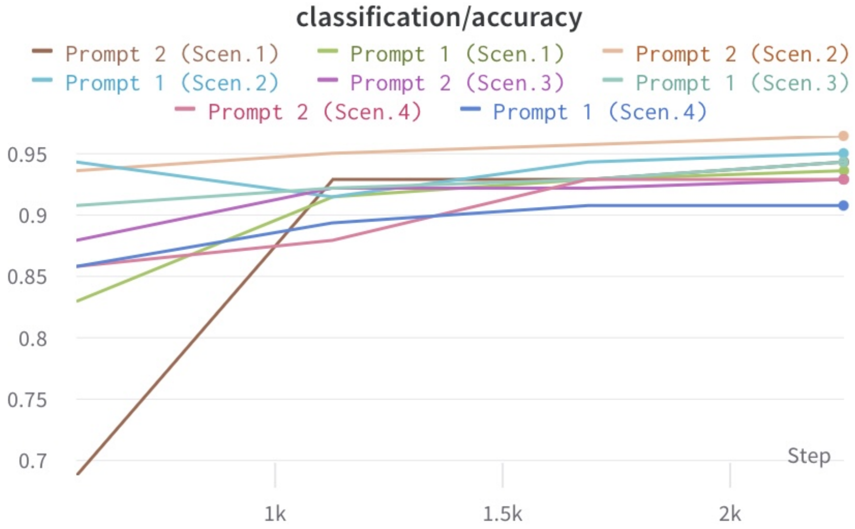


Figure 6: Accuracy for the four classification scenarios considering the two prompts.

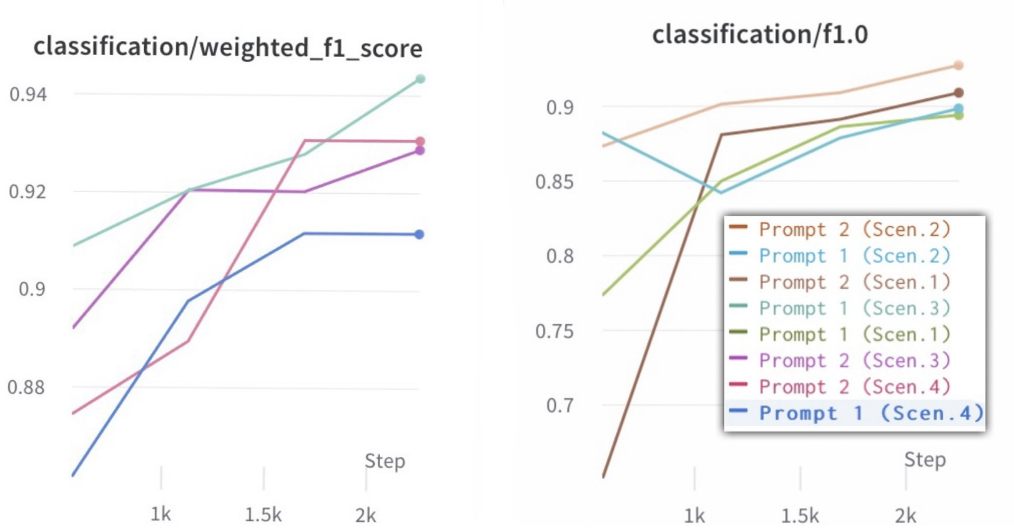


Figure 7: F1 scores for multiclass scenarios (scenarios 3 and 4) and for binary classifications (scenarios 1 and 2) on the left and right respectively.

In future work, we plan to test GPT-3 with a broader range of prompts in order to evaluate its performance in more complex and fine-grained tasks related to LRC and deontic modality classification. The potential to better interrogate the detail in this data may be particularly intriguing when combined with other NLP tasks like SRL (Humphreys et al., 2021). Another exciting prospect is creating expert systems that can identify legal requirements automatically and check compliance accordingly.

While the latest GPT-4 is currently unavailable for fine-tuning (OpenAI announced its upcoming accessibility), we plan to conduct a number of few-shot learning experiments with GPT-4 and compare the results with the current fine-tuning scenario. We also plan to employ other LLMs such as MPT(Team, 2023), Dolly-v2(Conover et al., 2023), BLOOM(Scao et al., 2022), LLaMA(Touvron et al., 2023).

8 Acknowledgements

Davide Liga was supported by the project INDIGO, which is financially supported by the NORFACE Joint Research Programme on Democratic Governance in a Turbulent Age and co-funded by AEI, AKA, DFG and FNR and the European Commission through Horizon 2020 under grant agreement No 822166.
Livio Robaldo has been supported by the Legal Innovation Lab Wales operation within Swansea University's Hillary Rodham Clinton School of Law. The operation has been part-funded by the European Regional Development Fund through the Welsh Government.

References

Ashley, Kevin D. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.

Athan, Tara, Harold Boley, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Wyner. 2013. Oasis legalruleml. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 3–12.

Boella, G., L. Di Caro, L. Humphreys, L. Robaldo, R. Rossi, and L. van der Torre. 2016. Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. *Artificial Intelligence and Law*, 4.

Boella, Guido, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. 2013. A system for classifying multi-label text into eurovoc. In *ICAIL*, pages 239–240. ACM.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.

2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chalkidis, Ilias, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical rnns. *arXiv preprint arXiv:1805.03871*.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Conover, Mike, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

de Maat, Emile and Radboud Winkels. 2008. Automatic classification of sentences in dutch laws. In *Legal Knowledge and Information Systems*. IOS Press, pages 207–216.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dragoni, Mauro, Serena Villata, Williams Rizzi, and Guido Governatori. 2018. Combining natural language processing approaches for rule extraction from legal documents. In Ugo Pagallo, Monica Palmirani, Pompeu Casanovas, Giovanni Sartor, and Serena Villata, editors, *AI Approaches to the Complexity of Legal Systems*, pages 287–300, Cham. Springer International Publishing.

Gao, Xibin and Munindar P Singh. 2014. Extracting normative relationships from business contracts. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 101–108.

Gomez-Perez, Jose Manuel, Ronald Denaux, and Andres Garcia-Silva, 2020. *Hybrid Natural Language Processing: An Introduction*, pages 3–6. Springer International Publishing, Cham.

Humphreys, Llio, Guido Boella, Leendert van der Torre, Livio Robaldo, Luigi Di Caro, Sepideh Ghanavati, and Robert Muthuri. 2021. Populating legal ontologies using semantic role labeling. *Artificial Intelligence and Law*, 29:171–211.

Joshi, Vivek, Preethu Rose Anish, and Smita Ghaisas. 2021. Domain adaptation for an automated classification of deontic modalities in software engineering contracts. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1275–1280.

Kiyavitskaya, Nadzeya, Nicola Zeni, Travis D Breaux, Annie I Antón, James R Cordy, Luisa Mich, and John Mylopoulos. 2008. Automating the extraction of rights and obligations for regulatory compliance. In *International Conference on Conceptual Modeling*, pages 154–168. Springer.

Liga, Davide. 2022. Hybrid artificial intelligence to extract patterns and rules from argumentative and legal texts.

Liga, Davide and Monica Palmirani. 2020. Transfer learning with sentence embeddings for argumentative evidence classification. *Proceedings of the 20th Workshop on Computational Models of Natural Argument*.

Liga, Davide and Monica Palmirani. 2022a. Deontic sentence classification using tree kernel classifiers. In *Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 1*, pages 54–73. Springer International Publishing Cham.

Liga, Davide and Monica Palmirani. 2022b. Transfer learning for deontic rule classification: the case study of the gdpr. In *INTERNATIONAL CONFERENCE ON LEGAL KNOWLEDGE AND INFORMATION SYSTEMS, Saarbrücken 14-16 December 2022*. EasyChair.

Neill, James O', Paul Buitelaar, Cecile Robin, and Leona O' Brien. 2017. Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 159–168.

Nguyen, Minh-Phuong, Thi-Thu-Trang Nguyen, Vu Tran, Ha-Thanh Nguyen, Le-Minh Nguyen, and Ken Satoh. 2022. Learning to map the GDPR to logic representation on DAPRECO-KB. In *ACIIDS (1)*, volume 13757 of *Lecture Notes in Computer Science*, pages 442–454. Springer.

Palmirani, Monica and Fabio Vitali. 2011. Akoma-ntoso for legal documents. In *Legislative XML for the semantic Web*. Springer, pages 75–100.

Radford, Alec and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Robaldo, L. 2021. Towards compliance checking in reified I/O logic via SHACL. In Juliano Maranhão and Adam Zachary Wyner, editors, *Proc. of 18th International Conference for Artificial Intelligence and Law (ICAIL)*. ACM.

Robaldo, L., S. Villata, A. Wyner, and M. Grabmair. 2019. Introduction for artificial intelligence and law: special issue ”natural language processing for legal texts”. *Artificial Intelligence and Law*, 27(2):113–115.

Robaldo, Livio, Cesare Bartolini, and Gabriele Lenzini. 2020. The dapreco knowledge base: representing the gdpr in legalruleml. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5688–5697.

Robaldo, Livio, Cesare Bartolini, Monica Palmirani, Arianna Rossi, Michele Martoni, and Gabriele Lenzini. 2020. Formalizing gdpr provisions in reified i/o logic: the dapreco knowledge base. *Journal of Logic, Language and Information*, 29(4):401–449.

Rodríguez-Doncel, Víctor, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor. 2020. Introduction: A hybrid regulatory framework and technical architecture for a human-centered and explainable ai. In *AI Approaches to the Complexity of Legal Systems XI-XII*. Springer, pages 1–11.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shaghaghian, Shohreh, Luna Yue Feng, Borna Jafarpour, and Nicolai Pogrebnyakov. 2020. Customizing contextualized language models for legal document reviews. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2139–2148. IEEE.

Song, Dezhao, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 106.

Sun, X. and L. Robaldo. 2017. On the complexity of input/output logic. *The Journal of Applied Logic*, 25:69–88.

Team, MosaicML NLP. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Waltl, Bernhard, Johannes Muhr, Ingo Glaser, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2017. Classifying legal norms with active machine learning. In *JURIX*, pages 11–20.

Wyner, Adam and Wim Peters. 2011. On rule extraction from regulations. In *Legal Knowledge and Information Systems*. IOS Press, pages 113–122.