



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA



HYPERMODELEX



European Research Council  
Established by the European Commission



Co-funded by  
the European Union



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

# LLM for the legislative domain

**Monica Palmirani, Michele Corazza**  
CIRSFID-ALMA AI, University of Bologna, Italy

AI in Legislative process could be considered High Risk

In case of use of Generative AI we should

“Generative foundation models, like GPT, would have to comply with additional transparency requirements, like **disclosing that the content was generated by AI**, designing the model to prevent it from generating illegal content and publishing summaries of copyrighted data used for training.” EU Parliament, June 2023

AT A GLANCE  
Plenary – June 2023



Parliament's negotiating position on the  
artificial intelligence act

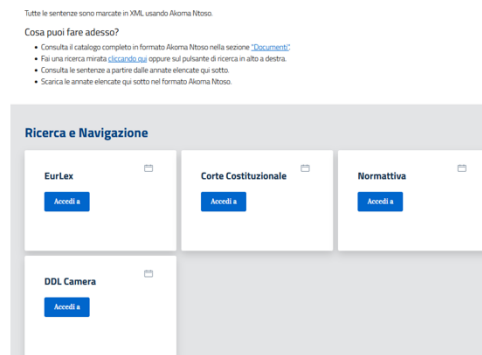


ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

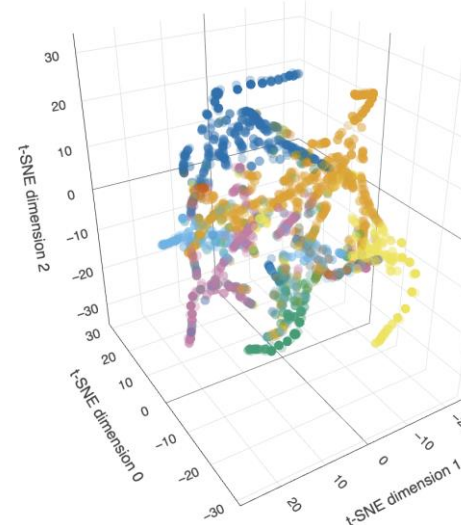
# Some scenarios

1. Suggest the relevant definitions according to title, topic, keywords (EUROVOC)
2. Suggest the relevant normative references from an incomplete prompt (partial citation)
3. Similarity between different legal sources (PDL, Corte Costituzionale/ amendments, Regolamenti EU)
4. Use LLM for extracting the temporal modifications
5. Extraction of «obligations/mandates/exception» and formalization in AKN-XML and RDF
6. Model plain legislation in AKN-XML
7. Generate «preamble» and «definitions» using AKN

## A. Ricerca avanzata



5. Naviga i documenti originali in EUR-LEX e Normativa *point-in-time*



1. Cercami le **definizioni** in EUR-LEX e in Normativa di «energy/energia» **affini** alla mia tematica

2. Cercami i **referimenti** in EUR-LEX **affini** al tema «hydrogen» usando EUROVOC

3. Cercami i **PDL simili** sul tema «energia» (secondo le classificazioni tematiche della Camera)

4. Cercami le decisioni della **Corte Costituzionale** in tema di «energia»



Camera dei deputati  
XIX LEGISLATURA  
Comitato di vigilanza sull'attività di documentazione

### Premiazione dei vincitori

della Manifestazione di interesse per la raccolta di proposte per l'utilizzo dell'intelligenza artificiale generativa per la Camera dei deputati

25 luglio 2024, ore 11:30  
Sala della Regina

Intervengono:  
**Lorenzo Fontana**  
Presidente della Camera dei deputati  
**Anna Ascani**  
Vicepresidente della Camera dei deputati

SEGUI LA DIRETTA:  
WEBTV.CAMERA.IT

Per accrediti: sg\_ufficiostampa@camera.it



# Critical issues in legal domain

**Structure:** LLM works at **sentence level/document level** and this approach is not capable to understand the structure (e.g., sequence of articles)

**Context:** LLM loses the **context** (e.g., jurisdiction, temporal parameters)

**Innovation:** LLM depends to the **past data series** (e.g., new brilliant solution has no historical series)

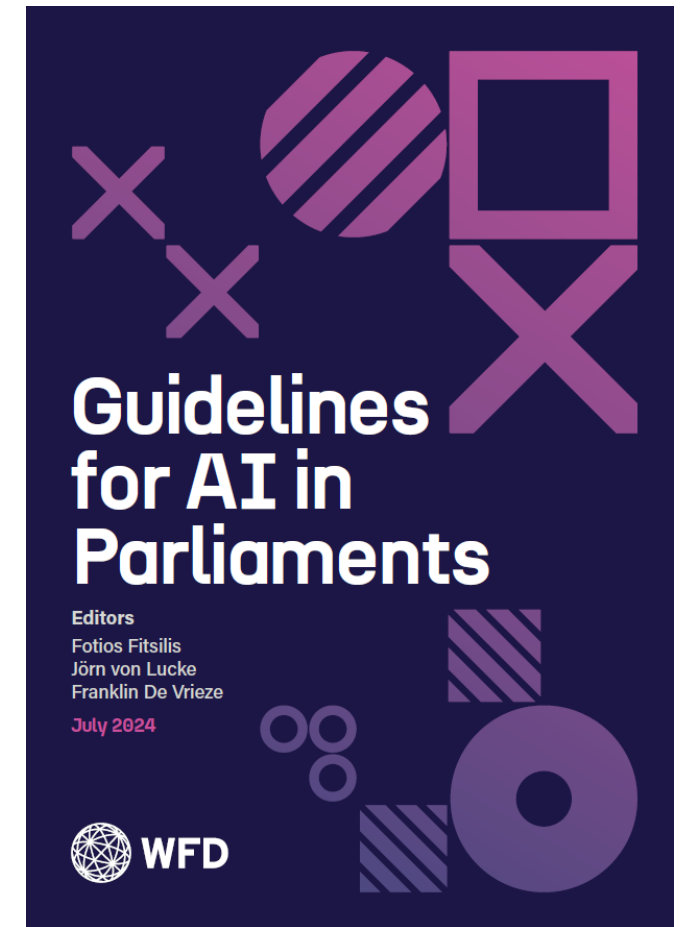
**Reference:** ML does not consider the **normative and juridical citations**. The normative references evolve over time (e.g., art. 3 is not the same forever)

**Time:** the LLM is **timeless** and the legislation is integrated in the legal system



# Critical issues in legal domain

- Provenance of the legal sources
- Data/Platform sovereignty
- Explicability, Transparency, Accountability
- Parliamentary Autonomy
- Separation of Power
- Integrity of democratic processes
- Free Mandate
- Continuity of Power



## Part 2.

# Guidelines

## for AI in parliaments



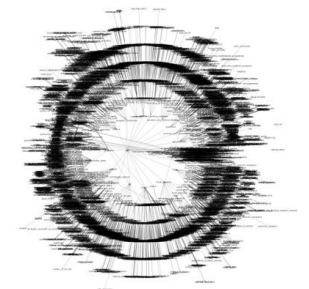
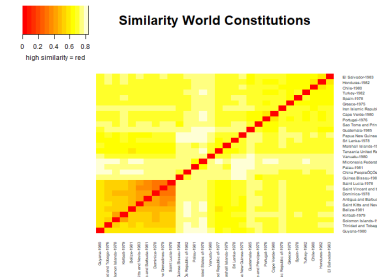
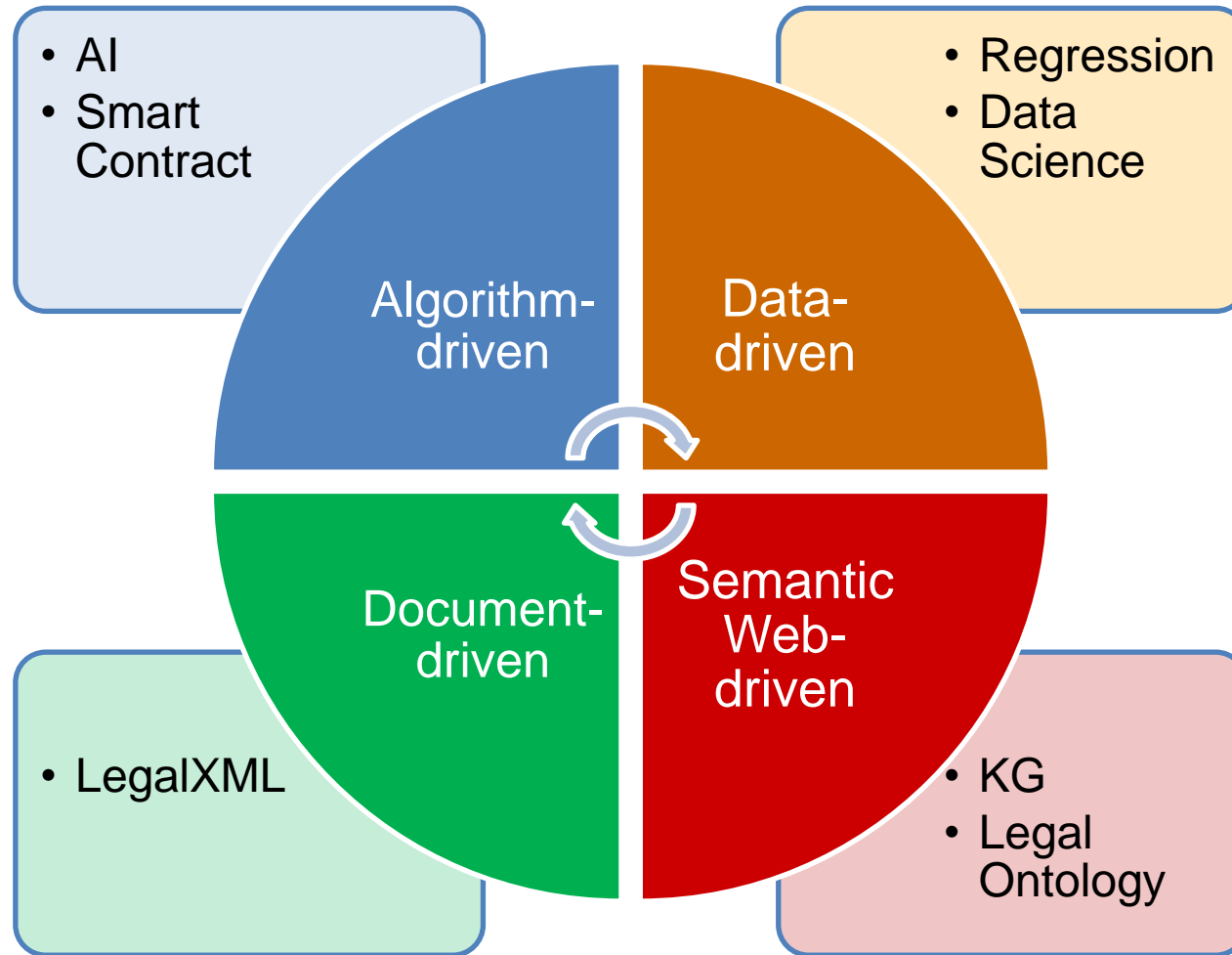
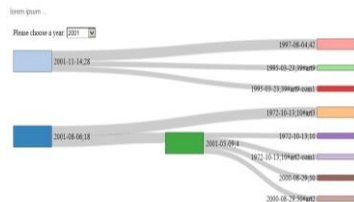
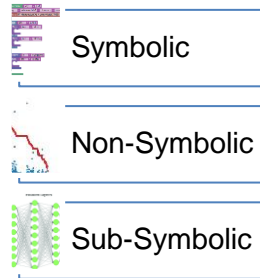
### Summary of the guidelines

<b>1. Ethical principles</b>	<b>24</b>	<b>2. Artificial general intelligence (AGI) and human autonomy</b>	<b>36</b>	<b>4. AI governance and oversight</b>	<b>54</b>	<b>5. AI system design and operation</b>	<b>62</b>
1.1. Accountability and transparency	26	2.1. Promotion of human autonomy	38	4.1. Integration into a broader digital parliamentary strategy	56	5.1. Implementing standardised data schemes and processes	64
1.2. Respect for human dignity, rights, and privacy	27	2.2. Ethical requirements for designers and developers	39	4.2. Efficient data governance and management protocols	57	5.2. Emphasising AI algorithms' explainability	65
1.3. Fairness, equity, and non-discrimination	28	2.3. Recognition of AGI as a real prospect	40	4.3. Establishing a parliamentary ethical oversight body	58	5.3. Building robust and reliable AI systems	66
1.4. Addressing biases in data and algorithms	29			4.4. Assessing the effects of parliamentary AI	59	5.4. Regulating the use and deployment of AI systems	67
1.5. Upholding intellectual property rights	30	<b>3. AI privacy and security</b>	<b>42</b>	4.5. Securing access to and control over the data	60	5.5. Assessing risk	68
1.6. Preservation of human values and cultural diversity	31	3.1. Embedding safety and robust security features	44	4.6. Cooperation with stakeholders	61	5.6. Monitoring and evaluating AI systems	69
1.7. Evaluation and mitigation of unintended consequences	32	3.2. Including privacy-by-design concepts	45			5.7. Agreeing minimum accuracy levels	70
1.8. Public participation and engagement	33	3.3. Secure processing of personally identifiable information	46				
1.9. Respect for the rule of law and democratic values	34	3.4. Outsourcing considerations	47			<b>6. AI capacity building and education</b>	<b>72</b>
1.10. Promotion of policy goals	35	3.5. Consideration of data sovereignty issues	48			6.1. Establishing expert teams	74
		3.6. Ensuring the integrity of source material	49			6.2. Organising training programmes	75
		3.7. Risk of overreliance on AI	50			6.3. Supporting knowledge exchange and cooperation	76
		3.8. Securing training and testing data	51			6.4. Documenting AI-related activities	77
		3.9. Human oversight in security decisions	52			6.5. Public education about the use and limits of AI in parliament	78



# Hybrid AI for the Legal Domain

Content, Context, Semantic, Processing



# Goals

- We aim to produce a LLM model for legislative domain in the context of EU documents
- This model should understand and produce Akoma Ntoso XML
- Our case study is based on the automatic generation of **preambles** or definitions



AKOMA NTOSO

Architecture for Knowledge-Oriented Management of African  
Normative Texts using Open Standards and Ontologies



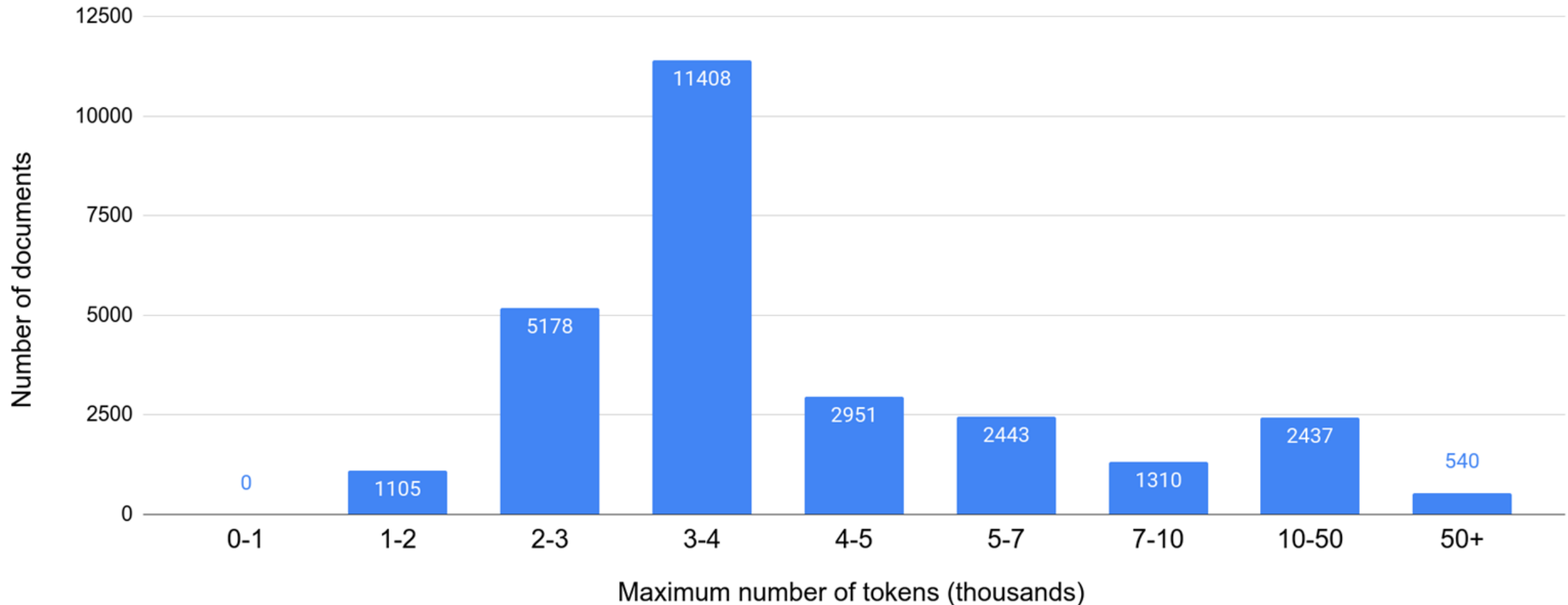


# The necessity for an encoder model

- To generate preambles, it is crucial to **retrieve the pertinent legislative documents that should be cited in the preamble**
- This is not optional, as the model should keep on working even when new documents, not seen during training, are approved
- The retrieval of these normative references is the goal of our **encoder model**



# The length of legislative documents



# Why not a pre-trained encoder?

- Legislative documents include sentences semantically connected (e.g., obligation – penalties, preamble – articles)
- Legislative documents contain normative references, which are sometimes crucial to understand the semantic content of a sentence, for example:  
*"(50) 'personal data' means personal data as defined in Article 4, point (1), of Regulation (EU) 2016/679;"*
- Our idea is to use a **hierarchical, reference-aware** model

## Article 5: Prohibited AI Practices

Date of entry into force: 2 February 2025  
According to: Article 113(a)  
Inherited from: Chapter II

[See here for a full implementation timeline.](#)

### SUMMARY +

1. The following AI practices shall be prohibited:

(a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; [Related: Recital 29](#)

(50) 'personal data' means personal data as defined in Article 4, point (1), of Regulation (EU) 2016/679;

(51) 'non-personal data' means data other than personal data as defined in Article 4, point (1), of Regulation (EU) 2016/679;



# Greedy splitting: optimizing the token usage

- The input to our hierarchical model, we first need to split the document following the structure of the AKN XML format
- To do this, we traverse the tree using a depth-first algorithm and we stop at the last element before we exceed the specified number of tokens (typically 512)
- The procedure is then repeated, starting after the end of the last examined element



# Greedy splitting: an example

```
<act name="EuropeanUnionRegulation">
  <meta>
    <identification source="#cirsfid">
      <FRBRWork>
        <FRBRthis value="/akn/eu/act/regulation/2009/822/!main" />
        <FRBRuri value="/akn/eu/act/regulation/2009/822/" />
        <FRBRalias value="32009R0822" name="CELEX" />
        <FRBRdate date="2009-08-27" name="Act Date" />
        <FRBRauthor href="#EP" />
        <FRBRcountry value="eu" />
        <FRBRnumber value="822" />
      </FRBRWork>
      <FRBRExpression>
        <FRBRthis value="/akn/eu/act/regulation/2009/822/eng@2009-08-27/!main" />
        <FRBRuri value="/akn/eu/act/regulation/2009/822/eng@2009-08-27/!main" />
        <FRBRdate date="2009-08-27" name="Act Date" />
        <FRBRauthor href="#EP" />
        <FRBRlanguage language="en-EN" />
      </FRBRExpression>
      <FRBRManifestation>
        <FRBRthis value="/akn/eu/act/regulation/2009/822/eng@2009-08-27/!main.xml" />
        <FRBRuri value="/akn/eu/act/regulation/2009/822/en-EN@2009-08-27.xml" />
        <FRBRdate date="2009-08-27" name="Act Date" />
        <FRBRauthor href="#EP" />
      </FRBRManifestation>
    </identification>
  </meta>
```

```
<preface>
  <longTitle>
    <p><inline name="uppercase">Commission Regulation</inline> (EC) No 822/2009</p>
    <p>of 27 August 2009</p>
    <p>amending <ref href="/akn/eu/act/regulation/2005/396/~annex_II">Annexes_II</ref>, <ref
      href="/akn/eu/act/regulation/2005/396/~annex_III"> III</ref> and <ref
      href="/akn/eu/act/regulation/2005/396/~annex_IV"> IV</ref> to <ref
      href="/akn/eu/act/regulation/2005/396/">Regulation (EC) No_396/2005</ref> of the
      European Parliament and of the Council as regards maximum residue levels for azoxystrobin,
      atrazine, chlormequat, cyprodinil, dithiocarbamates, fludioxonil, fluroxypyr, indoxacarb,
      mandipropamid, potassium tri-iodide, spirotetramat, tetraconazole, and thiram in or on
      certain products</p>
    <p>(Text with EEA relevance)</p>
  </longTitle>
</preface>
<preamble>
  <formula name="preambleFormula">
    <p>THE COMMISSION OF THE EUROPEAN COMMUNITIES,</p>
  </formula>
  <citations eId="cits_1">
    <citation eId="cits_1__cit_1">
      <p>Having regard to the <ref href="/akn/eu/act/treaty/1957/TCEE/eng@2009-08-27/!main">Treaty
        establishing the European Community</ref></p>
    </citation>
    <citation eId="cits_1__cit_2">
      <p>Having regard to <ref href="/akn/eu/act/regulation/2005/396/eng@2009-08-27/!main">Regulation
        (EC) No_396/2005</ref> of the European Parliament and of the Council of 23 February 2005
        on maximum residue levels of pesticides in or on food and feed of plant and animal
        origin and amending Council <ref
          href="/akn/eu/act/directive/1991/414/eng@2009-08-27/!main">Directive 91/414/EEC</ref><noteRef
          href="#f00001" class="footnote" marker="1" />, and in particular Article <ref
          href="/akn/eu/act/directive/1991/414/eng@2009-08-27/!main~art_5__para_1">5(1)</ref>
          and Article <ref href="/akn/eu/act/directive/1991/414/eng@2009-08-27/!main~art_14">14</ref></p>
    </citation>
  </citations>
  <recitals eId="recs_1">
    <intro eId="recs_1__intro_1">
      <p>Whereas:</p>
    </intro>
```



# How to represent references

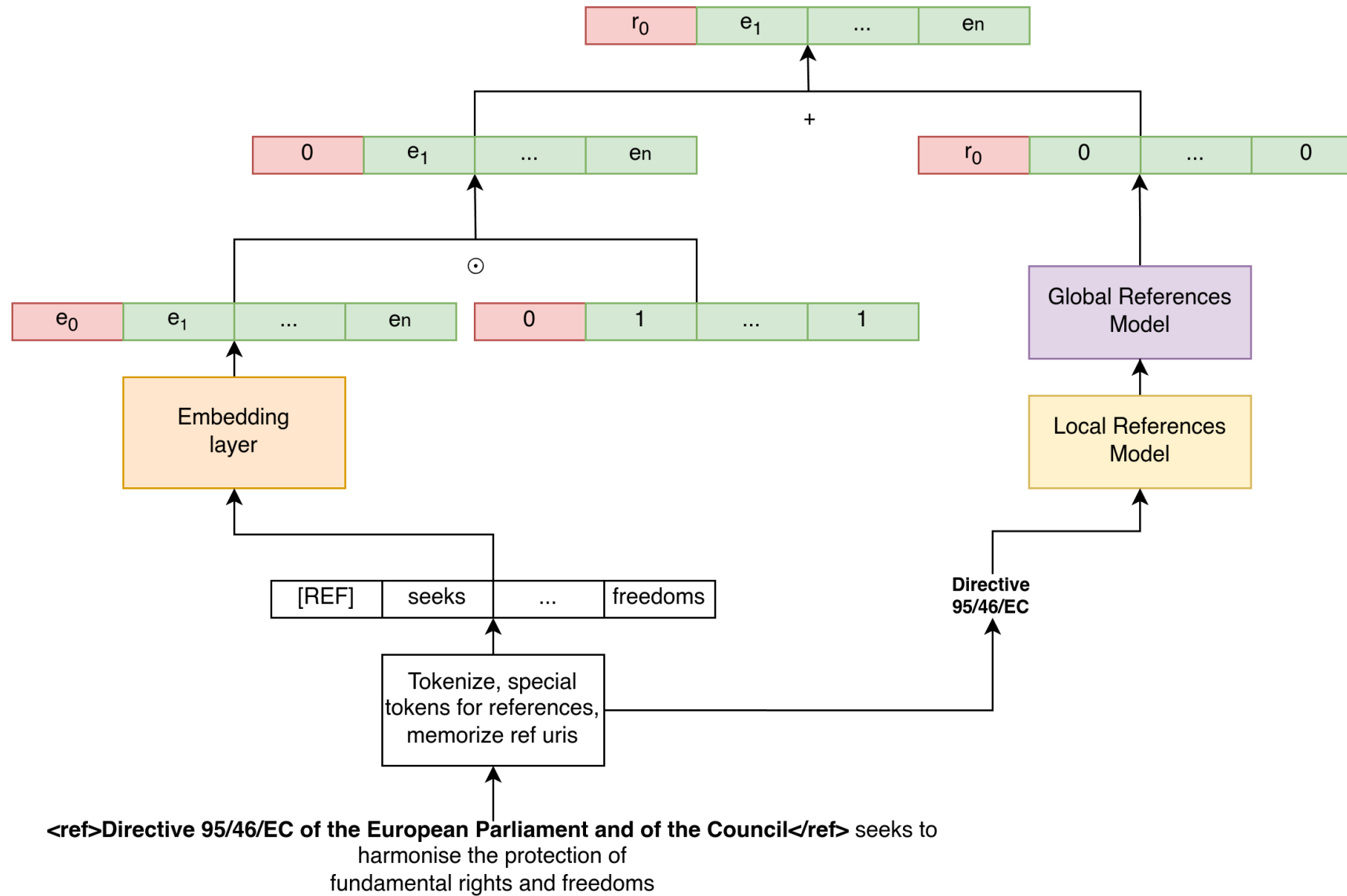
- For our goals, it is important to treat normative references as a special token (we will see why later);
- Luckily, AKN has a special element for references:

```
<p>The other members shall be appointed in accordance with <ref  
href="/akn/eu/act/regulation/2010/1093/!main">Regulation (EU) No  
1093/2010</ref> and <ref  
href="/akn/eu/act/regulation/2010/1095/!main">Regulation (EU) No 1095/2010</ref>  
</p>
```

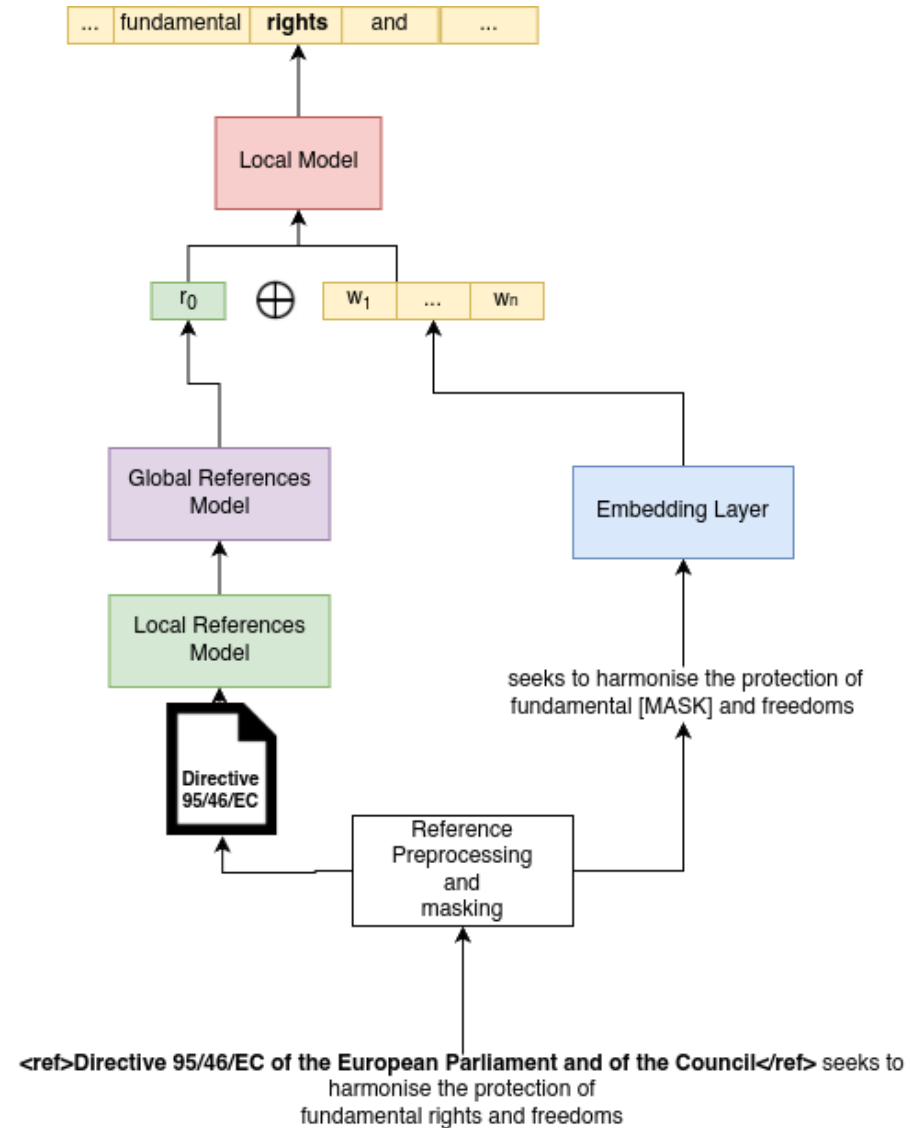
- Therefore, we can replace these elements and their content with the special token "[REF]";
- While doing so, we also memorize their **position in the sequence of tokens** and the URI associated with each.



# References masking and preprocessing

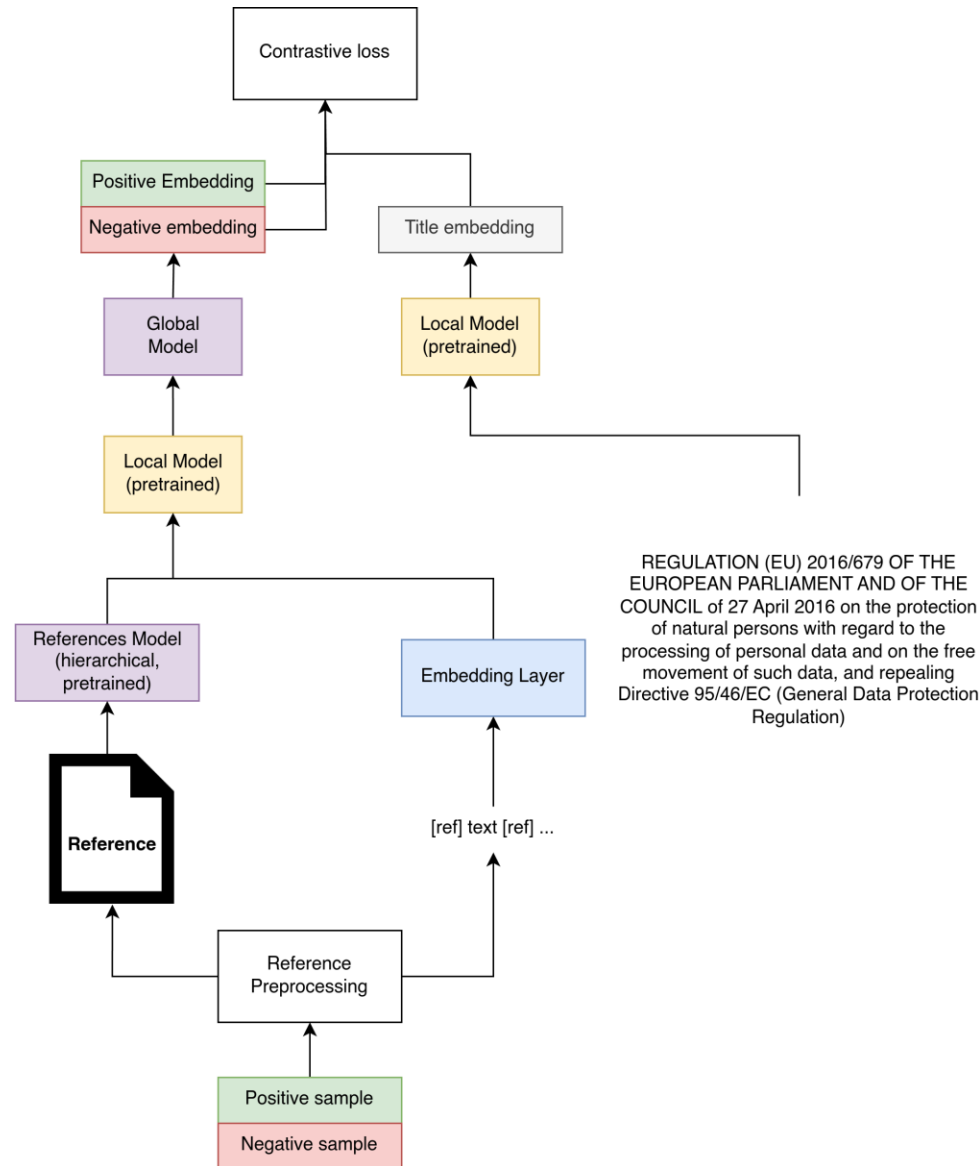


# Model training: MLM

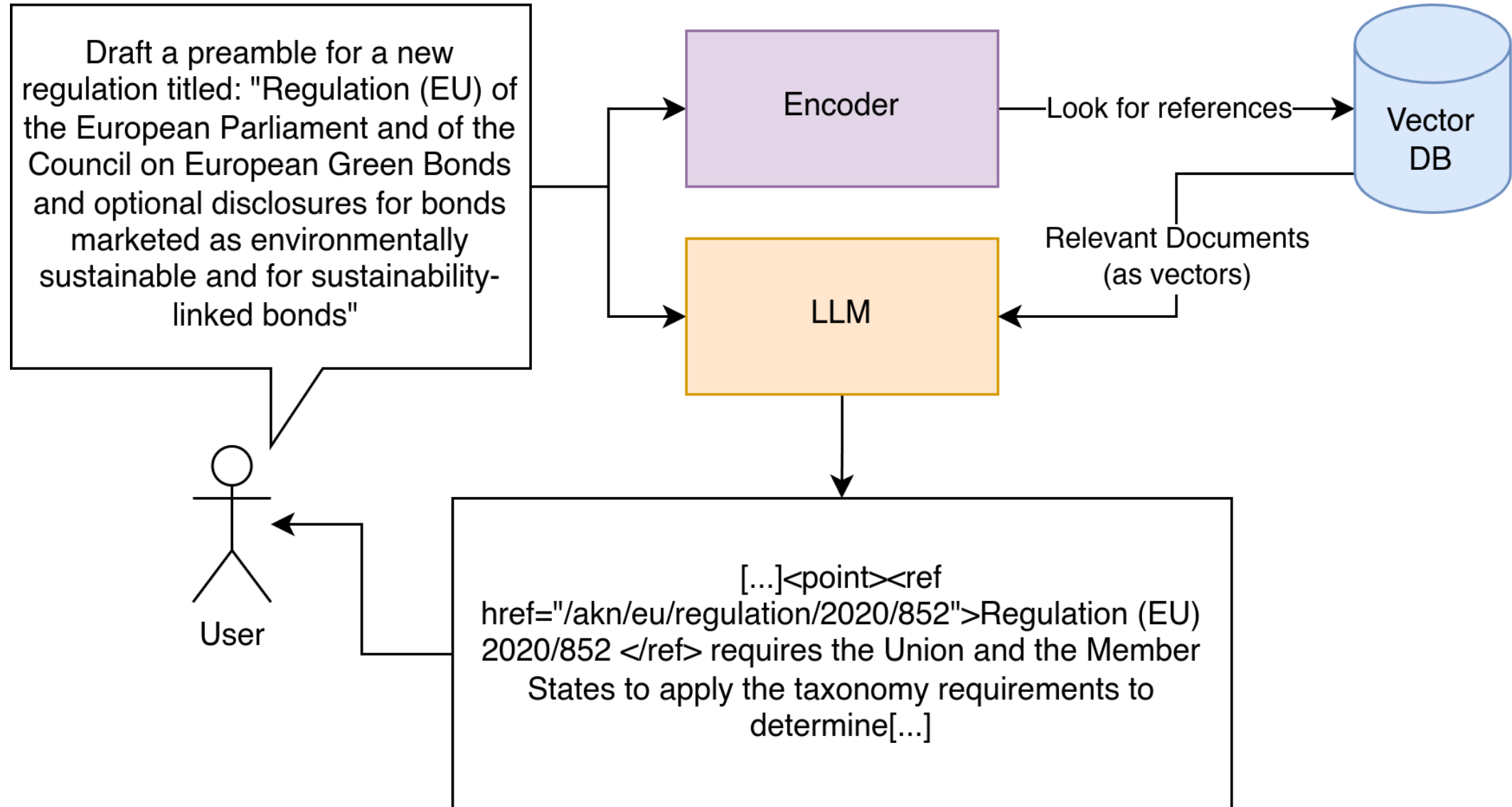




# Model: training for Information retrieval



# A vision of the future



# Disadvantages and doubts

- The hierarchical nature of the model allows us to process very long documents, but it is **very expensive** in terms of time and memory during training
- Using a pre-trained encoder does not allow us to run the model on a single A100
- Using a pre-trained tokenizer from a general-purpose model is not helpful when dealing with AKN XML documents (it uses a lot more tokens)
- Is a single embedding enough for an entire document? Perhaps this approach should be scaled, and the model should produce larger embeddings
- Do the gradients flow inside this very deep model, or do we have problems with **vanishing gradients**? We are asking a single token to influence the entire result



# The advantages of this model

- The model is **reference-aware**, meaning that it can understand the semantic relations between the legislative documents
- The model's hierarchical nature allows it to process  $512 * 512 = 260k$  tokens, as opposed to the 512 normally used for encoder models
- The outputs of the model allow us to build a **vector database** that can be used to retrieve the most relevant documents, we will use **re-ranking** as well
- Using **Akoma Ntoso** allows us to discard abrogated documents and to provide the system with the latest consolidated version of any document



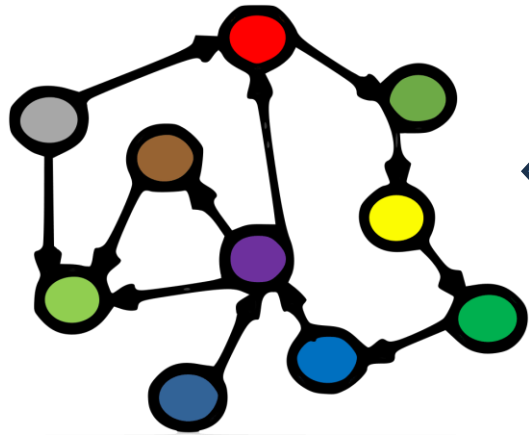
# Where we are

- We obtained a collection of approximately 40k legislative EU documents (3.8 GiB)
- We then proceeded to discard less relevant aspects of the document and metadata (tables, annexes, attachments, lifecycle, proprietary, restrictions, ...). The resulting dataset is measurably smaller (2.0 GiB)
- We developed the entire model, including the references resolution mechanism and splitting mechanism
- We are ironing out issues using a smaller dataset for question answering for now (Stanford Question Answering Dataset 2.0)



# FRAMENET

# KG4AKN



DL



**FrameNet:** D4  
D2

**Filtro RAG**  
(symbolic)

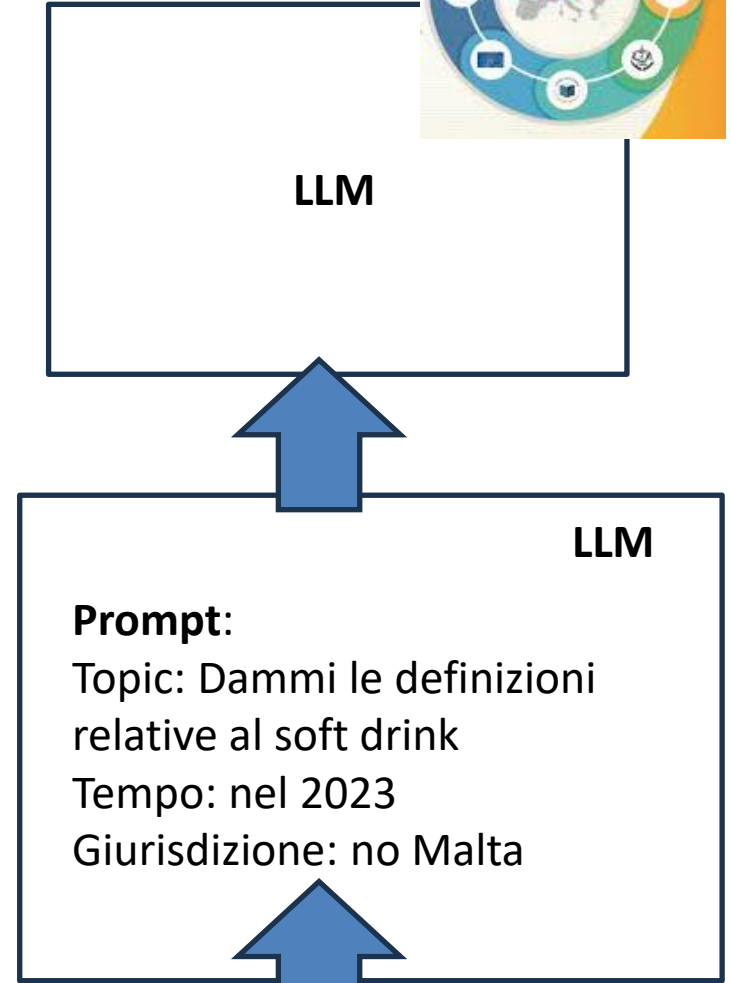
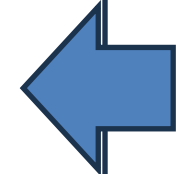
Spiegare perché no

- D1
- D3

Perché sì

- D4
- D2

D1  
D2  
D3



**LLM**

**Prompt:**  
Topic: Dammi le definizioni relative al soft drink  
Tempo: nel 2023  
Giurisdizione: no Malta

**LLM**

Natural Language

