ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LLM for Legal Explanations

Giuseppe Contissa

27 Settembre 2024

Three projects

- Different ways to use LLMs for legal explanation

- In 3 research projects

- **FACILEX**: To explain code in natural language
- **POLINE**: To explain legal decisions by extracting principles of law
- **DAFNE**: To explain platforms' DSA statement of reasons in light of ToSs

Research Team

Legal

- Dr. Piera Santin
- Dr. Alessia Fidelangeli

Legal-Informatics

- Dr. Federico Galli
- Dr. Marco Billi
- Dr. Andrea Filippo Ferraris

Computer Scientists

- Dr. Marco Aspromonte
- Dr. Galileo Sartor

FACILEX – LLM for Explainability (from code to text)

- Facilex: a Decision Support System in the context of mutual recognition instruments
- Knowledge Engineers formalize positive law (EU and National Acts) in a programming language (Prolog)
- Answers are provided in the programming language, thus LLMs are implemented for translating the code back into natural language
- The model should be able to extract from the Prolog language the following pieces of information:
  - a simplified summary of the norms relevant to the inference;
  - the list of rights granted to the user according to the input facts representing the case;
  - a description of the inference process that led the system to its solution.

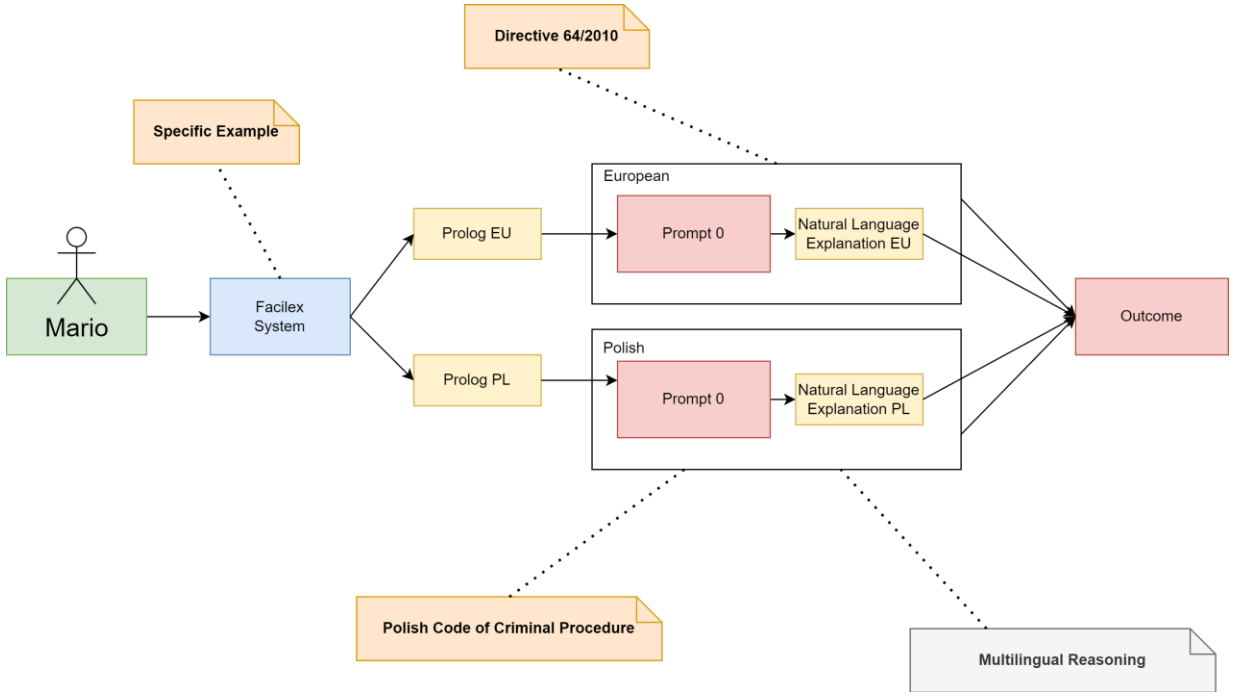
FACILEX – LLM for Explainability

What is needed for Prompt Engineering:

- Format of the language (how the laws have been formalized)

- Context of the domain (which laws have been applied)
- Outcome Expectations (what elements should be highlighted)

- Input: Articles from the EU and National Acts + their formalization in Prolog code
- Output: Natural Language Explanation of the code



FACILEX – LLM for Explainability

Excerpt of the answer provided by the Prolog system:

has_right(art3_1, mario, right_to_translation, essentialDocument):-
proceeding_language(mario, dutch) [FACT]
essential_document(art3_3, mario, documents)
authority_decision(mario, essential_document) [FACT]
not(person_understands(mario, dutch))

Excerpt of the outcome of the LLM:

Summary:[…]right to have essential documents translated […]

What Rights do You Have: Right to Translation of Essential Documents: […]

Why do You Have Them: Right to Translation of Essential Documents: This right is based on the fact that you do not understand the language of the proceedings (not(person_understands(mario, dutch))) and that there are documents considered essential for your defense (essential_document(art3_3, mario, documents)) as determined by the authority (authority_decision(mario, essential_document)).

POLINE – LLM for Explainability

- POLINE: retrieval and analysis of judicial principles of law (JPOL) in the context of Value Added Tax (VAT) case-law
- Legal analysis: what is a JPOL? How do we define it?
- LLMs implemented for automatic identification and extraction of JPOLs from the judgements

POLINE – LLM for knowledge extraction II

- Input: text of the judgement.
  - Preprocessing such as:

    removal of the preamble

    extraction of the motivation

    removal of the decision
  - Paragraph as the standard for classification

    1 JPOL = 1 paragraph

- Output:

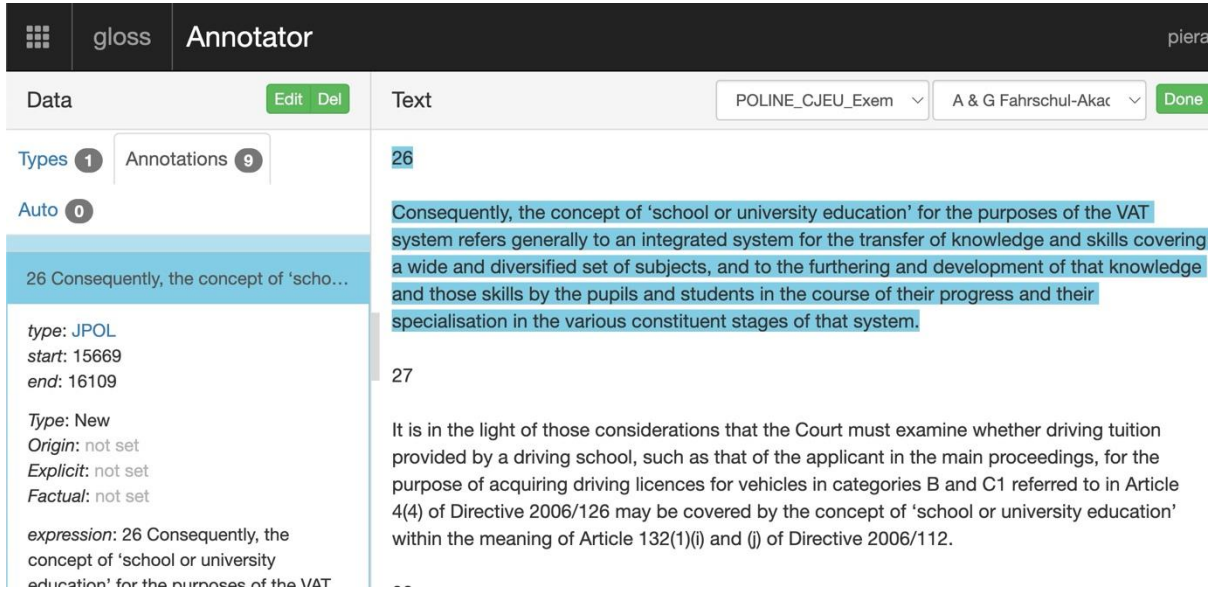
  A JPOL is a portion of text, extracted from the argumentative part of a judgement

POLINE – LLM for knowledge extraction III

- Prompt Engineering
  - Define what is JPOL:
    - Interpretation of a rule, of the portion of a rule, or of a general principle.
    - Consequences stemming from the interpretation/application of a rule or a principle in a legal system.
    - Subsumption of a fact within a rule.
    - Qualification of a factual hypothesis as a concept contained within a rule.
  - Define what is NOT a JPOL:
    - Not be a rephrase of the legislation.
    - Not recall of what the CJEU said in a previous paragraph of the same decision.

    Approaches: few-shot learning (provide examples taken from previous judgements) + paragraph classification (each paragraph should be autonomously classified as a JPOL)

Example of a JPOL and its Extraction

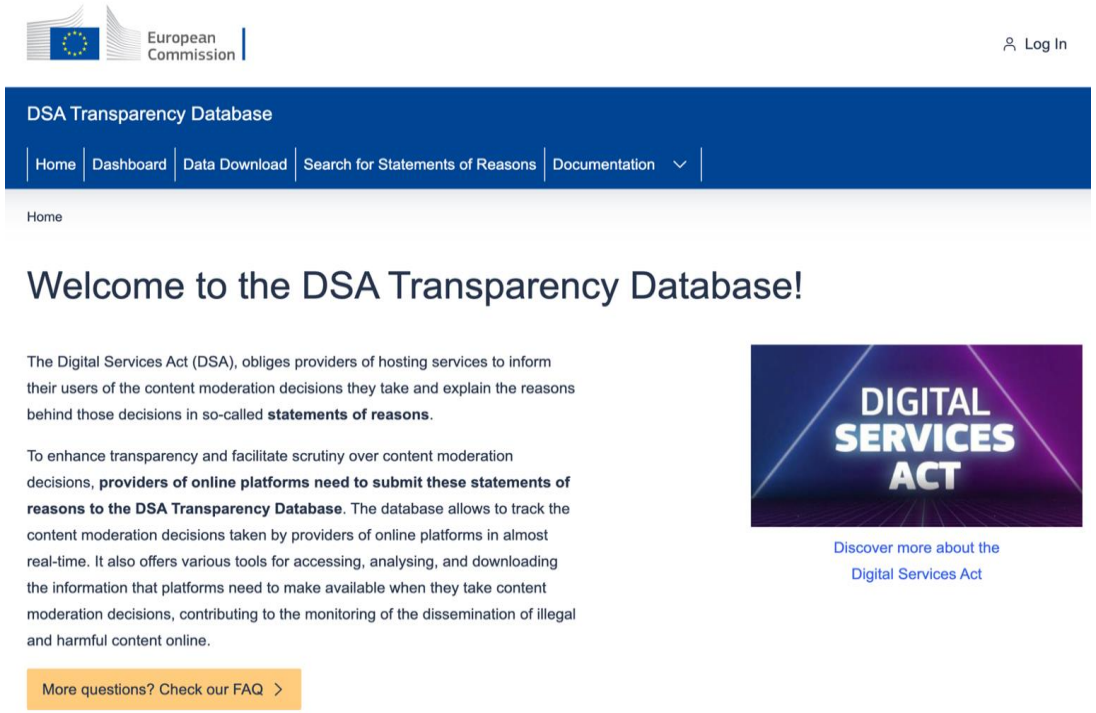



```
### Reasoning:

- **Paragraph 17** interprets the scope of Article 132 exemptions, indicating they are limited to certain public interest activities.
- **Paragraph 18** highlights the autonomous nature of these exemptions to ensure uniform application across Member States.
- **Paragraph 19** discusses the strict interpretation required for these exemptions, balancing it against their intended effect.
- **Paragraph 22** expands the concept of 'school or university education' beyond traditional qualifications to include broader educational activities, provided they are not purely recreational.
- **Paragraph 26** defines 'school or university education' as an integrated system for transferring a broad set of knowledge and skills.
- **Paragraph 30** applies the interpretation to conclude that driving tuition does not fall under 'school or university education' for VAT exemption purposes.

These paragraphs meet the criteria for JPOLs as they interpret legal concepts and principles rather than merely restating legislation or recalling previous decisions.
```

DAFNE – LLM for explaining Statement of Reasons

- **Context:** The Digital Services Act (DSA) requires hosting service providers to inform users about the content moderation actions they undertake and to provide explanations for these decisions = Statements of Reasons (SoRs).
- **Problem**: Platforms often provide vague, complex explanations when removing or restricting user content. Terms of Service (ToS) are frequently cited but can be difficult for users to understand.
- **Objective**: Use of Large Language Models (LLMs) to enhance the clarity of SoRs.
- **Solution**: Develop a multi-agent LLM system to link SoRs with relevant sections of the platform's ToS.

DAFNE – Data



Data Source: Custom dataset compiled from the **DSA Transparency Database**.
Platforms Analysed:

1. **Booking.com** - Commercial content moderation.
2. **Reddit** - User-generated, community-driven content.
3. **LinkedIn** - Professional networking and business communication.

Time Frame:
SoRs selected between March 2024 - August 2024

Dataset Composition:

- Total of **7,000 Statements of Reasons (SoRs)**:
  - **3,000 from Booking.com**
  - **2,000 from Reddit**
  - **2,000 from LinkedIn**

SoR Attributes Analysed:

- **UUID**: Unique identifier for each SoR.
- **Ground for Incompatible Content**: Reason for content violation.
- **Explanation**: Detailed reason for removal.
- **Decision Facts**: Facts supporting the content restriction.

DAFNE – Workflow

Vector Store Creation (Blue Box)

- **Purpose**: Prepare the platform's Terms of Service (ToS) for retrieval.
- **Process**: Chunk ToS into sections, convert to vectors using an embedding model (Voya

Retriever and Similarity (Red Box)

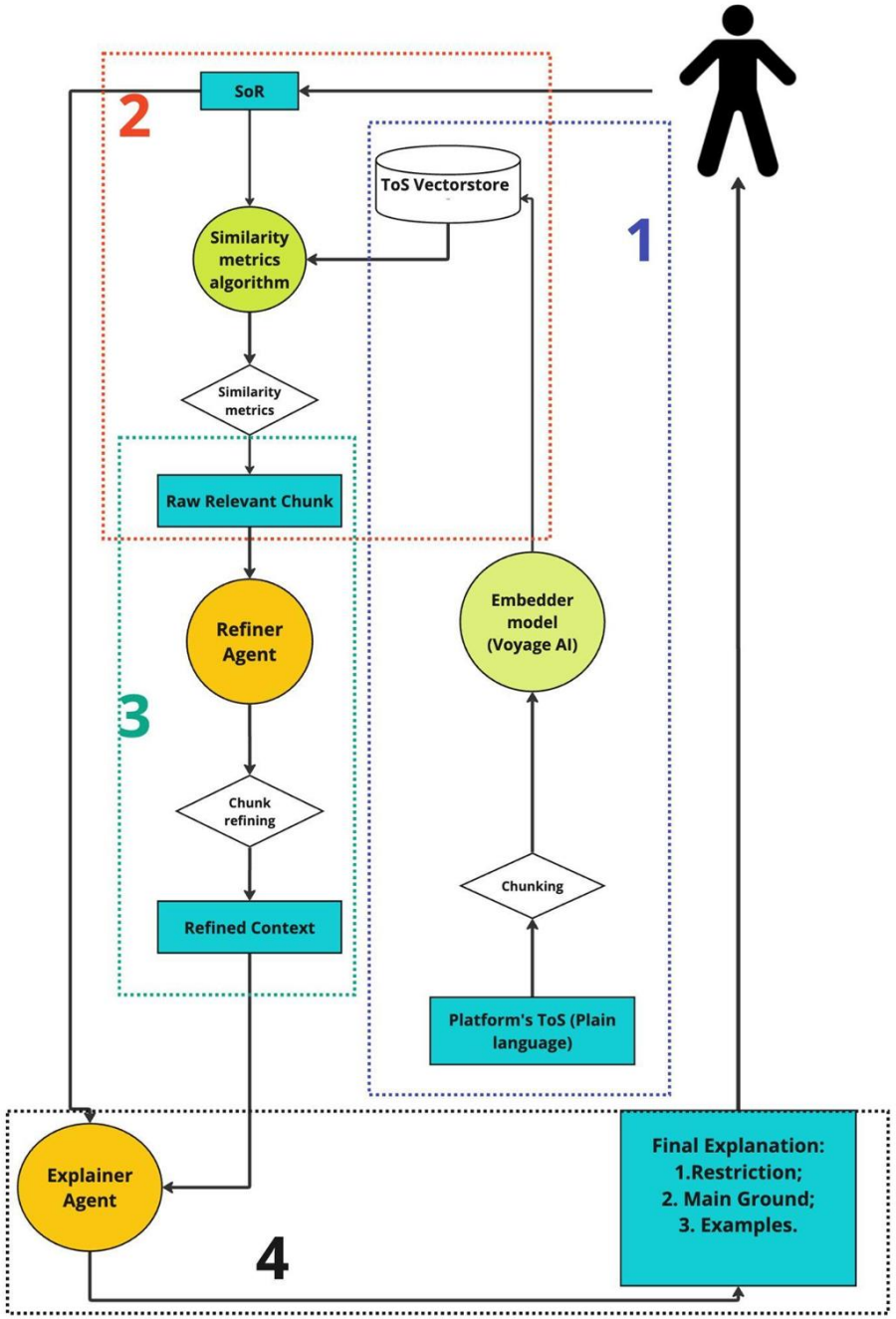
- **Purpose**: Retrieve relevant ToS sections for a given SoR.
- **Process**: Use a hybrid similarity approach (Cosine Similarity + BM25) to select the top **Chunk**.

Refiner Agent (Green Box)

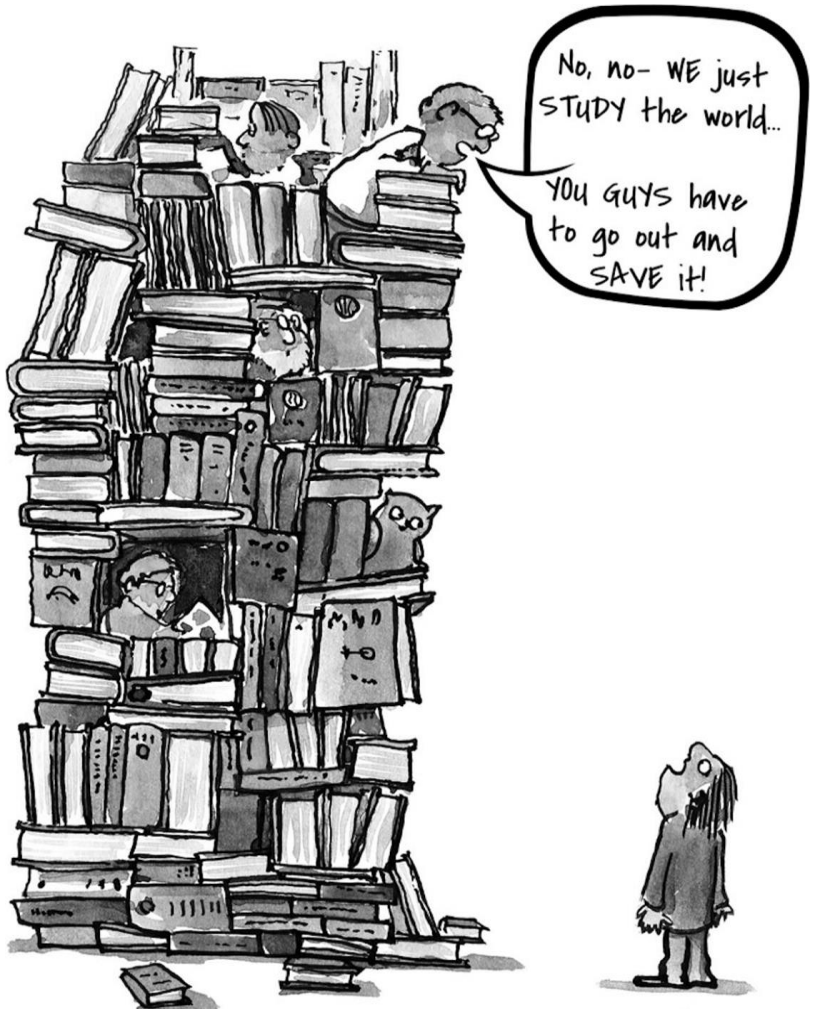
- **Purpose**: Refine the retrieved chunks for clarity.
- **Process**: Refiner Agent filters the Raw Relevant Chunk to remove irrelevant content, g

Explainer Agent (Black Box)

- **Purpose**: Provide a user-friendly explanation.
- **Process**: Explainer Agent uses the Refined Context to generate explanations, including
  1. **Restriction** – Action taken.
  2. **Main Ground** – Relevant ToS rule.
  3. **Examples** – Illustrations of rule application.



DAFNE – Validation



Evaluation Approach:

- **Human evaluation** chosen for nuanced feedback (Conducted by **three evaluators** with expertise in the field)
- Ensures assessment beyond automated metrics

Key Validation Metrics (Rated 1-5):

1. **Relevance**: Checks if the output aligns with the SoR and ToS. High scores mean strong relevance, low scores show misalignment.
2. **Accuracy**: Ensures key details from the ToS are retained. High scores mean complete retention, low scores indicate omissions.
3. **Coherence**: Evaluates if the output logically follows the ToS without adding unrelated content.
4. **Readability** (only for Explainer): Rates clarity and ease of understanding. High scores indicate clear, consistent, and smooth explanations.

DAFNE –Results

Model	Platform	Relevance	Accuracy	Coherence
GPT4o-mini	Booking.com	**4.69**	3.84	4.38
	Reddit	4.45	3.80	4.60
	LinkedIn	4.56	4.0	**4.68**
Mistral-7b	Booking.com	4.07	**4.28**	4.5
	Reddit	4.0	4.05	4.5
	LinkedIn	3.81	3.75	4.37

Table 2: Results for Refiner Agent across Different Platforms

Model	Platform	Relevance	Accuracy	Coherence	Readability
GPT4o-mini	Booking.com	**4.85**	4.57	**4.85**	4.71
	Reddit	4.71	4.12	4.62	**5.0**
	LinkedIn	4.0	4.4	4.7	4.8
Mistral-7b	Booking.com	4.71	**4.73**	4.14	4.9
	Reddit	4.7	3.8	4.2	4.73
	LinkedIn	4.75	4.0	4.12	4.62

Table 3: Results for Explainer Agent across Different Platforms

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thank you for your attention!

giuseppe.contissa@unibo.it