LightPrivacy: A RAG-Powered AI for eXplainable GDPR Compliance Assessment

Francesco Balassone[1], Antonio Boccarossa[1], Francesco Brunello[1], Vincenzo Luigi Bruno[1], and Salvatore Cangiano[1]

[1]Università degli Studi di Napoli Federico II
[1]{f.balassone, a.boccarossa, f.brunello, vincenzol.bruno, salva.cangiano}@studenti.unina.it

March 2025

Contents

1 Introduction

Every day, millions of users accept **terms of service** and **privacy policies** without reading their contents, often due to their **complexity** and **length**, thereby exposing themselves to potential **risks** related to the management of their personal data.

The General Data Protection Regulation ([20]) of the European Union establishes **strict rules** for the **processing**, **transparency**, and **consent** in the use of personal data, aiming to ensure individuals' control over their data and protect their fundamental rights. However, the legal language of the GDPR is often **difficult** for non-expert users to interpret, making it challenging to determine whether a privacy policy is truly compliant with the regulation.

This work proposes the design of an innovative AI system that **automatically assesses the compliance** of privacy policies with the GDPR, provides a clear **summary**, and **identifies potential risks**.

The solution is based on **LightRAG**, a retrieval-augmented generation structure organized in a graph form, and two large language models (**LLMs**): the first dedicated to **reasoning** and **response generation**, and the second focused on **evaluating** the quality of the response, paying attention to **consistency**, **faithfulness**, and **causality**. A key element of the system is **explainability**, which ensures the transparency and interpretability of the evaluations, supported by a robust **causal reasoning** approach that explicitly outlines cause-effect relationships between regulatory requirements and the information found in the analyzed policies.

This project aims to bridge the gap between the complexity of privacy regulations and their practical application, providing a reliable tool for the automatic verification of GDPR compliance.

2 Existing Solutions and Limitations

While AI can enhance the accessibility and efficiency of legal documents, ensuring **explainability** in these systems is crucial. Legal decisions have direct consequences on individuals' rights, making it essential for AI systems to provide transparent, understandable reasoning behind their outputs. However, many AI models function as "black boxes," lacking clear explanations for their decisions.

2.1 AI Systems in Legal Domain

The use of AI systems in the legal domain has gained significant traction due to their potential to improve efficiency and accessibility of complex legal documents, particularly privacy policies. Traditional legal language often poses a substantial barrier to user comprehension due to its length and complexity, leading to situations where users accept terms of service without fully understanding their implications. This lack of clarity can result in uninformed consent and exposure to potential data protection risks [13].

AI systems, especially those based on Retrieval-Augmented Generation (RAG) models, offer promising solutions to this problem. Unlike conventional deep learning models, which often function as black-box systems with limited explainability, RAG models provide enhanced transparency by incorporating retrieved external information into their responses. This open-book approach allows the AI to reference specific legal documents, ensuring that its outputs are well-grounded and factually accurate.

The importance of explainability in legal AI systems cannot be overstated. As legal decisions directly affect individuals' rights and obligations, the AI's reasoning must be clear, interpretable, and based on verifiable sources. To achieve this, explanations should adhere to four key principles:

- providing evidence and support for each output,
- offering user-friendly explanations,
- accurately reflecting the AI's decision-making process,
- and acknowledging the system's knowledge limits by refraining from outputting low-confidence results [16].

However, the implementation of AI in the legal sector also comes with challenges. Issues such as client confidentiality and data protection must be carefully managed to prevent misuse of sensitive information. Moreover, AI models are susceptible to the hallucination problem – the tendency to generate incorrect or fabricated outputs. In the context of privacy policy analysis, this could lead to misleading assessments of a policy's GDPR compliance.

To address these challenges, it is essential to evaluate the correctness and groundedness of AI-generated responses. Correctness ensures factual accuracy and relevance, while groundedness verifies that key factual claims are supported by appropriate legal references. Misgrounded or ungrounded responses — where

claims misinterpret sources or lack valid citations — undermine the reliability of the system's output. By focusing on these evaluation criteria, AI systems can provide more trustworthy and comprehensible legal analyses, ultimately empowering users to make informed decisions about their data privacy [12].

2.2 Explainability in LLM

Large Language Models (LLMs) are complex systems often regarded as "black boxes," making their internal workings difficult to interpret. Explainability aims to make these processes transparent, offering benefits for both end-users and researchers. For users, explainability builds trust in the system and helps them understand the model's capabilities and limitations. For researchers and developers, it facilitates the identification of biases, risk assessment, performance improvement, and debugging [7, 17, 22].

The techniques are divided into two paradigms:

- **The Traditional Fine-tuning Paradigm:** This paradigm focuses on explaining how Large Language Models (LLMs) make predictions, either for specific instances (*local explanation*) or for the model's overall behavior (*global explanation*) [24].
  - ***Local Explanation*** aims to clarify how a model generates predictions for a specific input. Key approaches include:
    - *Feature Attribution-Based Explanation:* Measures the relevance of input features (e.g., words, phrases) to the model's predictions. Examples include gradient-based methods or SHAP values.
    - *Attention-Based Explanation:* Analyzes attention weights to identify which parts of the input the model focuses on, providing insights into its decision-making process.
    - *Example-Based Explanations:* Explains model behavior by showing how outputs change with different inputs. This includes:
      - **Adversarial Examples:** Inputs designed to mislead the model.
      - **Counterfactual Explanations:** Modified inputs that change the model's output.
      - **Data Influence:** Identifies training examples that most influence a prediction.
    - *Natural Language Explanation:* Generates human-readable text to explain the model's decisions, often using annotated explanations during training.
  - ***Global Explanation*** provides insights into the model's internal workings and what it has learned overall. Key methods include:
    - *Probing-Based Explanations:* Investigates the linguistic knowledge captured by the model during pre-training, often using

probing classifiers or datasets tailored to specific properties (e.g., grammar).
    - *Parameter-Free Probing:* Evaluates model performance on datasets designed to test specific linguistic capabilities without requiring additional classifiers.
    - *Neuron Activation Explanation:* Examines individual neurons to understand their role in model performance or their association with specific linguistic features.
    - *Concept-Based Explanation:* Maps inputs to human-understandable concepts, explaining predictions in terms of high-level ideas rather than low-level features.
    - *Mechanistic Interpretability:* Studies the model's internal "circuits" by analyzing how individual neurons and their connections contribute to its functionality.

- **The Prompting Paradigm:** Focuses on explaining how Large Language Models (LLMs) perform tasks based on input prompts, without requiring additional training. This paradigm includes techniques like *In-Context Learning (ICL)* and *Chain-of-Thought (CoT) Prompting*, which aim to make the model's reasoning process more transparent and interpretable.
    - *Explaining In-context Learning (ICL):* In-Context Learning refers to the ability of an LLM to generalize a specific task based solely on the examples provided in the input. Explainability in this context focuses on understanding **how these examples influence the model's behavior**.
        - **Techniques Used:**
            - *Contrastive Demonstrations:* Altering the provided examples, such as reversing labels, modifying input text, or adding complementary explanations.
            - *Saliency Maps:* Highlighting which parts of the input most influence the model's predictions.
        - **Key Findings:**
            - In smaller models (e.g., GPT-2), reversing labels reduces the importance of keywords in the text. However, in larger models (e.g., InstructGPT), the opposite occurs: even if the label is reversed, the model can "unlearn" pre-existing associations and give more weight to the label provided in the examples. For instance, if the word *amazing* is repeatedly associated with negative sentiment, larger models adapt and treat it as a negative indicator, even though it is typically a positive term.
            - **Larger models can ignore *semantic priors*** (pre-existing knowledge from pre-training) and learn new input-label associations, even when these are contradictory. Smaller models rely more heavily on their semantic priors.

- *Explaining Chain-of-Thought (CoT) Prompting:* CoT Prompting allows LLMs to follow a logical sequence of steps to arrive at an answer, making their reasoning process more transparent.
  - **Techniques Used:**
    - Analysis of *saliency scores* of input tokens to understand how CoT influences the model's behavior.
  - **Key Findings:**
    - **CoT Prompting Improves Stability and Coherence:**
    - When using CoT, the saliency scores of question tokens become more stable, meaning the model focuses more consistently on the right parts of the input (e.g., keywords in the question) compared to standard prompting.
    - With standard prompting, the model might overemphasize less important words or vary its focus depending on the context. CoT, however, leads to more focused and consistent attention distribution, resulting in more reliable answers.
    - **Intermediate Reasoning Steps Encourage Symbolic Imitation:**
    - Introducing errors in intermediate CoT steps (e.g., incorrect calculations or statements) often leads the model to follow these incorrect steps.
    - This suggests that the model tends to *imitate the structure and symbols of the provided reasoning steps* rather than truly understanding or solving the problem. In essence, the model "copies" the form of reasoning more than it comprehends it.
- *Representation Engineering:* This approach studies and manipulates the internal representations of LLMs to understand how concepts and functions are encoded within the network. It consists of two main phases:
  - **Representation Reading:** Analyzes the activity of the model's neurons to identify how concepts and functions are represented. Techniques inspired by *neuroimaging* (e.g., linear artificial tomography scans) and tools like *linear probes* are used.
  - **Representation Control:** Manipulates these internal representations to influence the model's output. For example, modifying specific vectors can induce more truthful or false responses.

  Key Findings:

  - Models like LLaMA-13B learn **linear representations of abstract concepts** (e.g., space and time) and have specialized neurons, with accuracy improving in larger models.

* In summary, Representation Engineering offers advanced control over LLMs, but further research is needed to fully evaluate its effectiveness.

Traditional explainability techniques require significant computational resources, making them difficult to apply to large-scale LLMs. This has spurred innovation in techniques like *CoT Prompting* and *In-Context Learning*, which are better suited to the scale and complexity of modern LLMs. These methods aim to make models more interpretable and reliable, though challenges remain in fully understanding their internal reasoning processes.

2.3 Evaluation of Explanation in Prompting Paradigm

Recent research has highlighted significant challenges in evaluating explanations generated by large language models (LLMs) within prompting paradigms, particularly in terms of **plausibility** and **faithfulness** [8].

Plausibility measures how understandable and convincing an explanation is to humans, but studies show that optimizing for human preferences does not necessarily improve *counterfactual simulability*, meaning that explanations often fail to help users predict model behavior in alternative scenarios [**Chen2023d**].

Similarly, **faithfulness**—the degree to which explanations accurately reflect the model's true reasoning—remains problematic [19].

Research has demonstrated that *Chain-of-Thought (CoT) explanations* often fail to acknowledge biases introduced in inputs, leading to misleading justifications. Surprisingly, smaller models tend to produce more faithful explanations than larger ones [10]. To address these issues, researchers propose *decomposing questions into sub-questions*, which has shown promise in improving faithfulness without compromising performance. These findings suggest that current approaches to LLM-generated explanations require refinement to balance human interpretability with genuine insight into model decision-making.

3 Causal Reasoning in Large Language Models

Large language models (LLMs) have gained popularity thanks to Transformer architectures, allowing them to tackle many NLP tasks. However, it remains controversial whether they can accurately emulate human causal reasoning.

LLMs contribute to causal reasoning (CR) in two ways: (1) acting as causal reasoning engines via fine-tuning, prompt engineering, and external tools, and (2) supporting traditional methods by extracting causal knowledge to enhance analysis.

3.1 Fine-Tuning and Prompt-Based Methods

3.1.1 Advantages

Fine-tuning efficiently transfers learned knowledge to specialized models, reducing training time and injecting causal knowledge [11]. Prompt-based methods such as Chain of Thought (CoT) [21] and CausalCoT [9] enhance step-by-step causal reasoning.

3.1.2 Limitations

Fine-tuning faces challenges in collecting interventional data and accurately modeling complex causal relationships. Prompt-based methods depend on the model's intrinsic causal knowledge, which is often insufficient due to the correlational nature of LLMs [23]. Additionally, crafting effective prompts requires expertise [15].

3.2 External-Tool Integration

3.2.1 Advantages

Integrating structured knowledge bases and computational tools helps LLMs represent causal relationships more effectively [14].

3.2.2 Limitations

Effectiveness depends on the quality of external resources, which may introduce errors. These tools are often domain-specific, limiting generalizability.

3.3 Causal-Consistency Chain-of-Thought

A promising approach is **CausalGPT** [18], which solves knowledge-based reasoning problems using a **multi-agent system** with a **Faithful Reasoner** and a **Causal Evaluator**.

3.3.1 Faithful Reasoner

Mimics human reasoning by following a causal step-by-step approach: understanding the problem, retrieving relevant knowledge, and solving subproblems sequentially.

3.3.2 Causal Evaluator

Verifies **causal consistency** through:

- **Non-causal perspective**: Checks logical correctness of each reasoning step.
- **Counterfactual evaluation**: Introduces alternative premises to detect contradictions.

3.4 Comparison of Reasoning Structures

- **CoT (Chain of Thought)**: Single agent reasoning, lacks verification.
- **Self-consistency CoT**: Multiple agents generate independent reasoning chains, selecting the most frequent answer.
- **CaCo-CoT (Causal-Consistency CoT)**: Includes an evaluator agent verifying causal consistency via non-causal and counterfactual assessments.

3.5 Evaluation of Causality

Evaluating how causal an LLM's reasoning is can be a complex process. In fact, it has been shown that the **CoT paradigm does not always consistently improve performance** [3, 1] and does not necessarily represent the true "reasoning process" within the system [10, 19].

In the literature, the evaluation of causality is always associated with measuring the accuracy of a model in answering questions that require causal reasoning to solve the problem. In fact, what is being estimated is the so-called "causal inference," which refers to the process of estimating the causal effect of an intervention or treatment on a specific outcome. While causal discovery focuses on identifying causal relationships, causal inference focuses on quantifying these relationships.

A recent study aimed to answer the question of **how and when CoT enables an LLM to reason causally, like a human, and when it does not** [2]. To achieve this, a **Structural Causal Model (SCM)** was constructed, dividing the problem into three components: **instruction (Z)**, **reasoning steps (X)**, and **response (Y)**.

The study examined and discussed the causal relationships among these variables, identifying four possible SCM scenarios:

- **Causal Chain**: The instruction generates **causal reasoning**, which then leads to a response, generally improved by the reasoning process.

- **Common Cause**: The response beliefs **already exist** in the model before reasoning takes place; both the reasoning process and the response are **caused solely by the instruction**.
- **Full Connection**: A **mix of the first two cases**, where reasoning partially influences the response but is also affected by pre-existing beliefs.
- **Isolation**: An **extreme case** where the instruction generates reasoning, but the response **depends on neither** the instruction nor the reasoning. This scenario is **not** addressed in the study.

Only the **first case** leads the model to respond correctly and faithfully. The **second and third cases** can create **unexpected spurious connections** between the instruction and the response, potentially causing **inconsistencies and an unfaithful response**.

Going into more detail, the problem can be modeled as a **causal analysis study**:

- The three random variables **Z**, **X**, and **Y** have been defined.
- Suppose that the **SCM entails a distribution** $P_{X,Y}$ with N_X and $N_Y \overset{\text{iid}}{\sim} N(0,1)$.
- We **intervene on X** to modify the distribution of **Y**.

$$P_Y^{do(X)} = P(Y|do(X))$$

The final objective is to determine whether the treatment causes an actual change in Y. To this end, the **Average Treatment Effect (ATE)** is defined, which represents the effect of an intervention. Specifically, it compares the distribution of Y with and without the treatment:

$$ATE = E(Y \mid do(X)) - E(Y)$$

3.6 In-Context Learning in the Evaluation

Since the main problem has been divided into three parts (instruction, reasoning, and response), it is necessary to introduce the concept of **In-context Learning** in more detail, as it will be a crucial aspect in evaluating the causal consistency of the model. Tom B. Brown et al. (2020) demonstrated how LLMs could "learn" without retraining the model, solely from the context provided in the prompt.

These strategies are now used for sensitivity analysis in testing. In particular, injecting reasoning into the context of the LLM system is very often used in the research domain.

- **Zero-shot prompting**: The system has a CoT injected but no examples.

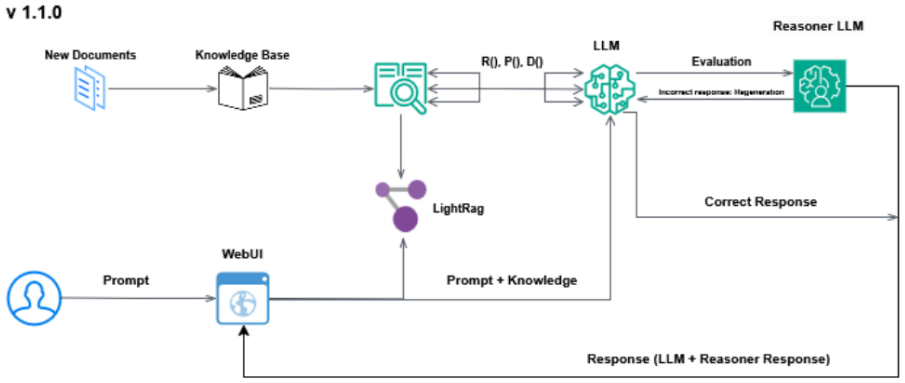
- **One-shot prompting**: The system has one example of a CoT and its associated response.
- **Few-shot prompting**: The system has multiple examples of CoT and their associated responses.

It is important to define these strategies because they are widely used as an additional parameter in the evaluation of LLM correctness. The choice of one strategy over another can significantly impact the model's response.

4 Proposed Solution

4.1 Architecture

Let us recall our goal: we aim to create a tool that can assist users in summarizing and verifying whether the terms and conditions of websites and services comply with the GDPR. The decisions and engineering process behind the development of the tool have been documented in another document (link, if available); below, we will instead outline the architectural solution adopted and the results obtained through a small testing campaign. The architecture is as follows:



The user will input instructions through a web page created with **Streamlit**. The system will then send the prompt to the framework used, **LightRag** [6], an environment that combines the simple RAG (Retrieval-Augmented Generation) approach with a newer approach called **GraphRAG**. The model, leveraging this framework, will create a graph of entities and relationships using the data available in the knowledge base. This graph, along with the data in the knowledge base, will be consulted to respond to the user's query.

Following the example of **CaCo-CoT**, our model has been built using a **multi-agent approach**. Specifically, there is a **single agent** that responds to the user's query. However, the response is not only provided to the client but is also evaluated by a **second agent**, which acts as the model's **evaluator**. The core idea is that the **evaluator**, using a **CoT approach**, assesses the correctness of the first agent's response. If the response **passes the evaluator's check**, the execution ends. Otherwise, the evaluator prompts the **reasoning agent** to generate a new response.

4.2 What is a GraphRag?

Retrieval-Augmented Generation (RAG) is a technology designed to accurately interpret a query and retrieve relevant information to incorporate into the response [4]. This approach leverages **vectorization** of both the query and a knowledge base (KB), measures the **similarity** between the query vector and KB vectors, and uses a **ranking mechanism** to extract the most relevant information. RAG was developed to **reduce hallucinations** typical of LLMs and to **"replace" fine-tuning** when resources and time for training are limited. Although this approach helps the model in the challenging task of being **explainable**, it struggles in certain situations. For example, **RAG has difficulty connecting multiple points** within a knowledge base and **understanding holistic knowledge**, making it less effective when dealing with **large documents and extensive KBs**.

To address these issues, Microsoft has explored a new approach that extends the classic RAG by integrating **knowledge graphs** [6]. The primary purpose of knowledge graphs is to uncover hidden relationships across different heterogeneous data repositories. This process was traditionally time-consuming and manual but highly valuable for various tasks. With the advent of LLMs and recent studies, knowledge graphs have become an extremely useful tool for enhancing the capabilities of RAG-based LLM systems. Specifically, the knowledge graph serves as:

- A **data store** for information retrieval.
- A **semantic structure** for extracting vector chunks.

Traditional RAG systems search for relationships and similarities in a text database. **GraphRAG** enhances this approach by incorporating the graph into the database. It not only searches for entities similar to the user's query but also retrieves related entities and descriptions, thereby adding more contextually relevant information.

The advantages of **GraphRAG** are:

- **Structured Representation**: Unlike vectors in a database, GraphRAG uses graphs composed of nodes (entities) and edges (relationships) to create a network that captures the complexity of contextual dependencies.
- **Contextual Knowledge**: It is ideal for recognizing triple relationships such as subject-predicate-object.
- **Explainability**: It improves the model's transparency by showing the relationships and entities used, helping users gain more trust in the system.

4.3 Framework Used: LightRag

To integrate this new technology into the system, and due to resource limitations, we adopted a lighter framework compared to GraphRAG, namely **LightRag** [6]. LightRAG improves information retrieval by creating an indexed

graph-based structure, rather than relying solely on raw textual documents. This allows for:

- **Segmenting information** into smaller, more manageable chunks.
- **Quickly identifying relevant data** without needing to analyze the entire text each time.
- **Extracting key entities and relationships** between concepts such as people, places, and actions.
- **Building a knowledge graph** that highlights the connections between these entities.

The indexed structure is represented as follows:

$$\hat{D} = (V, E) = Dedup \circ Prof(V', E'), \qquad V_i, E_i = \cup_{D_i \in D} Recog(D_i)$$

4.4 Architecture and Functions of the Graph-Based System

Let us define the core components involved in the graph-based indexing and evaluation process:

- **V** = graph nodes (the **entities extracted** from the text, such as names of people, places, or key concepts).
- **E** = graph edges (the **relationships between entities**, such as "a cardiologist treats heart diseases").
- **Recog(Di)** = **recognition function** that applies an LLM to **extract entities and relationships** from raw textual documents D_i.
- **Prof(V',E')** = **profiling function** that adds metadata to entities and relationships, *creating an optimized structure*.
- **Dedup** = **deduplication function** that removes redundancies by merging identical entities or duplicate relationships.

The indexing process begins with the extraction of entities and relationships. The LLM analyzes the documents to identify key entities and the relationships between them. These become the nodes and edges of the graph, a process made possible by segmenting the knowledge base into multiple chunks. Next, entity profiling takes place, where key-value pairs (K, V) are created. These metadata enhance search efficiency by allowing direct access to key information without needing to analyze the entire text. A graph cleaning phase is also performed, where duplicate entities or relationships are identified and merged, keeping the graph more compact and reducing the computational load for graph-based searches.

4.5 Consistency Evaluator: A Multi-Agent Approach

The explainability and causal abilities of foundation models are still a highly debated topic, but it is undeniable that they have greatly improved with methods such as integrating causal knowledge into training data and in-context learning techniques. **CaCo-CoT** aims to automatically integrate the latter approach. Our integration does not fully replicate the framework as presented in the study but instead implements a simplified version. In fact, the first prototype of the system utilized only a single "faithful reasoner" agent and a single "evaluator" agent. Despite the limitations of this approach—such as the perpetuation of cognitive biases when both the reasoner and the evaluator rely on incorrect conclusions—it still serves as a valuable support system that, if validated, can significantly enhance the model's explainability and its causal consistency.

5 Method

5.1 Evaluation of the System

To evaluate the accuracy of our system, we selected one of the tasks from **LegalBench**, a benchmark comprising various legal reasoning tasks [**legalbench**]. Specifically, we chose the **OPP-115** (Online Privacy Policies) dataset [**swilson˙acl˙2016**], a collection of website privacy policies annotated with details on data handling practices mentioned in the texts.

Our evaluation is framed as a binary classification task, where the system must determine whether a given policy clause satisfies an annotation intent, such as: "Does the clause describe how user information is protected?".

The categories of OPP-115 clauses consist of ten data practice categories:

- **First Party Collection/Use**: how and why a service provider collects user information.
- **Third Party Sharing/Collection**: how user information may be shared with or collected by third parties.
- **User Choice/Control**: choices and control options available to users.
- **User Access, Edit, & Deletion**: if and how users may access, edit, or delete their information.
- **Data Retention**: how long user information is stored.
- **Data Security**: how user information is protected.
- **Policy Change**: if and how users will be informed about changes to the privacy policy.
- **Do Not Track**: if and how Do Not Track signals for online tracking and advertising are honored.
- **International & Specific Audiences**: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
- **Other**: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

To bridge NLP research on privacy policies with the GDPR, Poplavska et al. [**poplavska2020**] have introduced a mapping between GDPR provisions and the OPP-115 annotation scheme. This connection has enabled automated classification of privacy policy text, demonstrating that the annotation scheme's underlying assumptions align with many of the topics mandated by the GDPR.

5.2 Dataset Sampling

The dataset was sampled based on the number of chapters associated with each category, normalizing the weights to balance the distribution of instances in the classification task. As shown in the figure, categories with a higher number of chapters (e.g., First Party Collection/Use and User Choice/Control) have a greater weight.

Column1	Column2	Column3	Columr	Columr	Columr
OPP-115 Category	GDPR Articles	GDPR Principles	#Capitoli	Pesi	#Istanze
First Party Collection/Use	4, 5, 6, 7, 8, 9, 10, 11, 24, 25, 30, 33, 34, 35, 36, 37, 38, 39, 89, 91, 95	1(a), 1(b), 1(c)	21	0,238636	9
Third Party Sharing/Collection	4, 6, 9, 19, 28, 29, 30, 37, 38, 39, 44, 45, 46, 47, 48, 49, 96	1(a), 1(b), 1(c)	17	0,193182	7
User Choice/Control	4, 6, 7, 8, 9, 13, 14, 17, 18, 20, 21, 26, 49, 77, 78, 79, 80, 82	1(a)	18	0,204545	8
User Access, Edit and Deletion	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 25	1(a), 1(d)	13	0,147727	6
Data Retention	5, 25, 30	1(e)	3	0,034091	2
Data Security	4, 5, 6, 12, 24, 25, 28, 30, 32, 33, 34, 35, 36, 45, 89	1(f)	15	0,170455	7
Policy Change	None	1(a)	0	0	
Do Not Track	None	None	0	0	
International and Specific Audiences	8	1(a)	1	0,011364	1
Other	None	None	0	0	

Figure 1: Dataset Sampling

Once the number of sentences was determined, a **random sampling** of the dataset was performed for each category, ensuring a fair distribution of samples across different privacy policy topics. This approach maintains consistency in the dataset while ensuring adequate representation of each category in the classification task.

5.3 Evaluation

To assess correctness, we evaluated **accuracy** across these classification tasks and compared results using different models for the **Reasoner agent**: **GPT-4o** and **Llama 3.1 (8B)**. By running the same classification queries across both models, we measured how effectively each model identified the correct category for a given policy clause.

Our evaluation provides insights into **how different LLMs interpret privacy policies**, highlighting potential advantages of larger, more advanced models in legal NLP tasks while also considering efficiency trade-offs for smaller-scale models like Llama 3.1 (8B).

We conducted experiments in two settings: **Zero-shot learning** and **Few-shot learning**. In the Zero-shot setting, models classified policy clauses without prior examples, relying solely on their pre-trained knowledge. In the Few-shot setting, models were provided with a small number of labeled examples to guide their predictions.

Category (sentences)	Llama3.1:8b Accuracy	GPT-4o Accuracy
First Party Collection/Use (9)	66,67%	44%
Third Party Sharing/Collection (7)	57,14%	57,1%
User Choice/Control (8)	25%	50%
User Access, Edit, and Deletion (7)	14,28%	42,86%
Data Retention (2)	50%	50%
Data Security (8)	37.5%	50%
International and Specific Audiences (2)	50%	50%
Total Accuracy	**41,86%**	**48,74%**

Category (sentences)	Llama3.1:8b Accuracy	GPT-4o Accuracy
First Party Collection/Use (9)	33,33%	33%
Third Party Sharing/Collection (7)	57,14%	57,1%
User Choice/Control (8)	75%	62,5%
User Access, Edit, and Deletion (7)	57,14%	57,1%
Data Retention (2)	100%	50%
Data Security (8)	50%	50%
International and Specific Audiences (2)	50%	50%
Total Accuracy	**55,81%**	**51,14%**

The results indicate that the overall accuracy of both models is not excellent, reflecting the complexity and specificity of the classification task, which differs significantly from our initial objective. However, the Few-shot learning approach improves performance by providing context through labeled examples, leading to better model predictions. This suggests that additional fine-tuning or more extensive examples could further enhance accuracy in future work.

6 Causal Analysis

Experiment	**GPT 4o Accuracy**	**Avg ‖ATE‖**
Zero-Shot CoT (Baseline)	0.90	-
Test: CoT (X) causes the Answer (Y) given a constant Instruction (Z)?		
Controlled (w/ default setting)	0.90	-
Treated (w/ golden CoT)	1	0.10
Treated (w/ random CoT)	0.60	0.40
CoT → Answer	**T**	0.25
Test: Instruction (Z) causes the Answer (Y) given a constant CoT (X)?		
Controlled (w/ default CoT)	0.90	-
Treated (w/ random instruction)	1	0.10
Treated (w/ random bias)	0.80	0.10
Controlled (w/ golden CoT)	1	0.10
Treated (w/ random instruction)	0.90	0
Treated (w/ random bias)	0.80	0.10
Instruction → Answer	**F**	0.08
Implied SCM Type	**I**	

In conclusion, our experiment confirms that the model adheres to a **Type I - Causal Chain SCM** ($Z \to X \to Y$), where the response (Y) is entirely determined by the reasoning process (X), rather than being directly influenced by the initial instruction (Z). This demonstrates an ideal causal reasoning structure, ensuring a step-by-step approach rather than an immediate jump to the final answer.

The integration of **LightRAG** further enhances this causal structure by reinforcing consistency in the Chain of Thought (X) and improving the separation between the instruction (Z) and the reasoning process. This leads to **greater coherence** in responses and minimizes the risk of direct influence from Z. However, a potential drawback is the **introduction of bias**, as any inaccuracies or limitations in the knowledge graph could distort the reasoning and shift the model towards a **Type III - Full Connection** SCM, where Y is influenced by both X and external information sources.

7 Future Works and Conclusions

Although this approach has achieved a certain level of explainability for the system, some limitations have been identified, particularly:

- Updating the knowledge base
- Inference time for responses

We therefore propose some ideas to expand the study and the overall project, which will be explored in future work.

7.1 Multi-agent Oracle

The main issue we encountered from the beginning was the need for a **comprehensive knowledge base** to make a foundation model more suitable for our **legal purposes**. As previously discussed, even in the **legal domain**, building a specialized agent for a specific task typically involves **fine-tuning**, which is usually **supervised**. The **lack of data** for our specific task and the **limited computational resources** available for fine-tuning led us to explore the **GraphRAG** approach. However, there is still a need for a **dataset** to expand the **knowledge base** beyond the **GDPR**.

One idea hypothesized during development is to extend the **multi-agent approach** by introducing a **pseudo-oracle**. This would consist of multiple **evaluator agents**, which, in addition to serving as a useful technique for **continuous online monitoring**, could **artificially expand the knowledge base**. This expansion would be achieved by incorporating the **reasoner's conclusions**, provided they are **validated by a majority of the evaluators**.

7.2 Chain of Draft

One of the main **limitations** of the system arises from its **multi-agent nature**: ideally, the user must wait for the **reasoner's response** to be generated and subsequently evaluated by the **evaluator**. This process may result in an **unacceptable response time**, further exacerbated by the **verbosity** of reasoning models during their inference process.

A very recent **innovative approach** has attempted to **constrain** the reasoning of these **LLMs** by limiting the **number of tokens** used in their reasoning process, obtaining promising results in terms of **trade-off** between the **number of tokens** allocated for reasoning and the **accuracy** of the generated responses [5].

References

[1] A. Sprague et al. “CoT paradigm does not always consistently improve performance”. In: *Proceedings of the 2023 International Conference on Machine Learning (ICML 2023)* (2023).

[2] G. Bao et al. “How and when CoT enables an LLM to reason causally, like a human, and when it does not”. In: *Journal of Machine Learning Research* 46 (2025), pp. 330–350.

[3] K. Kojima et al. “CoT paradigm does not always consistently improve performance”. In: *Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS 2022)* (2022).

[4] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *arXiv preprint arXiv:2005.11401* (2020). URL: https://arxiv.org/abs/2005.11401.

[5] Silei Xu et al. “Chain of Draft: Thinking Faster by Writing Less”. In: *arXiv preprint arXiv:2502.18600v2* (Mar. 2025). URL: https://arxiv.org/abs/2502.18600.

[6] Zirui Guo et al. “LightRag: A Lightweight Retrieval-Augmented Generation Framework”. In: *arXiv preprint arXiv:2410.05779* (2024). URL: https://doi.org/10.48550/arXiv.2410.05779.

[7] J. Bastings et al. “Explainability in AI”. In: *Journal of AI* 1 (2022), pp. 100–110.

[8] A. Golovneva et al. “Challenges in Evaluating Explanations from Large Language Models”. In: *arXiv preprint* 2212.07919 (2022). DOI: 10.48550/arXiv.2212.07919. URL: https://doi.org/10.48550/arXiv.2212.07919.

[9] Z. Jin. “CausalCoT: Enhancing causal reasoning in LLMs using Chain of Thought prompting”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML 2023)* (2023).

[10] N. Lanham et al. “Investigating Model Size and Faithfulness in Language Model Explanations”. In: *arXiv preprint* 2307.13702 (2023). URL: https://arxiv.org/abs/2307.13702.

[11] Y. Li. “Fine-tuning language models for causal reasoning”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2021)* (2021).

[12] Varun Magesh et al. *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. 2024. arXiv: 2405.20362 [cs.CL]. URL: https://arxiv.org/abs/2405.20362.

[13] János Papp. “Navigating the Digital Contract Maze: How AI can guide users to informed decisions”. In: *Magyar Nyelvőr* 148 (Jan. 2024), pp. 717–727. DOI: 10.38143/Nyr.2024.5.717.

[14] T. Pawlowski. "Integrating external knowledge bases for enhanced causal reasoning in LLMs". In: *AI & Knowledge Engineering Journal* 44.2 (2023), pp. 155–167.

[15] L. Perez. "Prompt engineering for large language models: A review". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.8 (2021), pp. 3047–3061.

[16] K.M. Richmond, S.M. Muddamsetty, and T. et al. Gammeltoft-Hansen. "Explainable AI and Law: An Evidential Survey". In: *DISO* 3 (1 2024). DOI: https://doi.org/10.1007/s44206-023-00081-z.

[17] H. Strobelt et al. "Interpretable models for large language models". In: *Journal of AI Research* 7 (2018), pp. 200–210.

[18] M. Tang. "CausalGPT: A multi-agent system for causal reasoning in language models". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)* (2025).

[19] T. Turpin et al. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track*. 2023.

[20] European Union. *General Data Protection Regulation*. Accessed: 2025-03-10. 2016. URL: https://gdpr-info.eu/.

[21] C. Wei. "Chain of Thought (CoT) prompting for causal reasoning in large language models". In: *Journal of Artificial Intelligence Research* 72.1 (2023), pp. 100–112.

[22] T. Yuksekgonul et al. "Improving Interpretability of Large Language Models". In: *AI Research* 3 (2023), pp. 300–310.

[23] L. Zhang. "Limitations of prompt-based causal reasoning in large language models". In: *Journal of Machine Learning* 58.3 (2024), pp. 410–428.

[24] Haiyan Zhao et al. "Explainability for Large Language Models: A Survey". In: *ACM Computing Surveys* 15.2 (2024). DOI: 10.1145/3639372. URL: https://doi.org/10.1145/3639372.