LLM-assisted Extraction of Regulatory Requirements: A Case Study on the GDPR

Sallam Abualhaija, Marcello Ceci, Nicolas Sannier, Domenico Bianculli,
Salomé Lannier, Martina Siclari, Olivier Voordeckers, Stanisław Tosza
University of Luxembourg, Luxembourg
Email: firstname.lastname@uni.lu

***Abstract*—Modern software systems increasingly rely on personal data. Despite the enforcement of the European General Data Protection Regulation (GDPR) and the growing awareness about privacy and data protection, many individuals' rights remain unsatisfactorily implemented in software systems. This is partially due to the knowledge gap between legal interpretation and software development.**

In this paper, we address this gap first by extracting, in close collaboration with legal experts, a list of 108 requirements pertinent to the right of access (`ACC`) and the right to portability (`PRT`), two fundamental rights under the GDPR. We further propose the XTRAREG approach, which utilizes large language models (LLMs) and retrieval augmented generation (RAG) to provide automated assistance in extracting privacy requirements from predefined legal sources.

Compared to the manually extracted requirements, XTRAREG can automatically generate requirements with an accuracy of 81.8% for `ACC` and 85.7% for `PRT`. Our empirical evaluation reveals two notable observations: (i) A skewed performance in terms of coverage in the favor of `ACC`, indicating the significant impact of abundant training data of the LLM, (ii) despite explicit exposure of legal references through `RAG`, the LLM generates requirements predominantly from the GDPR.

***Index Terms*—Privacy Requirements, General Data Protection Regulation (GDPR), Requirements Generation, Natural Language Processing (NLP), Large Language Models (LLMs), Retrieval Augmented Generation (RAG).**

I. Introduction

Modern software systems are becoming more complex and reliant on personal (often sensitive) data in return for providing more effective, intelligent, and personalized services. Artificial intelligence (AI)-enabled software systems are currently being integrated in a wide variety of application domains, directly interacting with—and affecting—humans [1]. For instance, conversational AI systems and modern chatbots are emergent technologies that have become widely used over a short time [2], [3]. These systems heavily depend on personal data.

Spurred by this technological advancement is the demand for new regulations and stronger compliance mechanisms to address the growing privacy concerns and ensure that personal data remains protected [4]. The *right to data protection* is a fundamental right established in a broad range of legal sources, such as the EU Charter of Fundamental Rights [5]. In 2016, the EU issued the General Data Protection Regulation (GDPR) [6], which provides a common legal framework on personal data protection for all EU member states. One of the objectives of the GDPR is to give individuals (data subjects) control over their personal data. In particular, Chapter III of the GDPR provides individuals with the rights to (1) access, rectify, or erase their personal data, (2) restrict or object to its processing, (3) ensure data portability, and (4) not be subject to decisions based solely on automated processing.

Supporting the development of lawful, ethical, and responsible software in this AI-driven transformation era is a central topic of requirements engineering (RE) and software engineering (SE) at large [7], [8], [9], [10]. However, extracting legal requirements is a challenging task for requirements engineers. Existing research explores innovative approaches for eliciting and representing legal requirements [11], [12], assessing the completeness of regulated documents (e.g., privacy policies) [13], [14], and ensuring the compliance of various software artifacts [15], [16].

A recent systematic mapping reports that GDPR has gained more attention in the RE literature over the past few years, with a majority of the work being conducted within the requirements elicitation phase. The study highlights, however, that 57 out of the 90 analyzed research papers have generic mentions of GDPR, while only 11 papers (out of 90) specifically investigate data subject rights [17, Table 12]. While the GDPR emphasizes individuals' rights, it does not provide the detailed guidance required for seamless integration into the software development process. Consequently, the compliance rate among EU companies remains low, with a recent survey indicating that 90% of data access requests are not fully answered on time [18]. From an RE standpoint, this lack of compliance can be partially attributed to the challenge of extracting and implementing a comprehensive list of legal requirements, pertinent to the data subject rights in this case. Such an activity involves not only selecting the most suitable legal sources for the application context, but it also requires parsing documents full of legalese to extract the relevant technical details.

To illustrate, consider the following example on the *right to portability*. Article (Art.) 20 of the GDPR states: *"The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller..."*. Taking into account that a data subject is the user of a given software system, this legal statement

Using GDPR only
PRT01. Upon a generic portability request, the system shall transfer to the user all data provided by them.

Using GDPR + Additional Sources
PRT02. The data transfer shall include all personal data that a data controller has collected upon consent or according to a contract, e.g., through location data, cookies, preferences, fitness data, purchase history, or credit card number.

Fig. 1: Example Requirements for the Right to Portability

can be translated into the requirement PRT01, outlined at the top of Fig. 1. While the above requirement captures the right to portability, the GDPR remains vague regarding certain implementation details, including how the request should be made (e.g., through a dedicated request form or a generic email), what exactly should be transferred, the time limit for the transfer, or the format in which personal data should be transferred. Leaving such details open for interpretation during the software development phase can lead to a shallow implementation of the legal requirement, potentially resulting in the development of non-compliant software due to lack of understanding of this right.

Additional legal sources may provide details that could improve the understanding of the GDPR statement listed above. For example, the *Guidelines on the right to data portability* [19] define, inter alia, personal data concerning a data subject as: *"Data actively and knowingly provided by the data subject (for example, mailing address, user name, age, etc.) [and] observed data provided by the data subject by virtue of the use of the service or the device. They may for example include a person's search history, traffic data and location data. It may also include other raw data such as the heartbeat tracked by a wearable device."* Capturing this detailed view of personal data in a second requirement (PRT02 at the bottom of of Fig. 1) clearly shows that data portability concerns not only data explicitly provided by the user but also data collected with the user's consent.

The above example illustrates that relying solely on the GDPR for achieving compliance might not be sufficient. Considering additional sources leads to capturing details necessary for appropriate implementation of legal requirements. The importance of leveraging complementary sources beyond the GDPR becomes evident in compliance activities such as audits. Instead of potentially ensuring de facto compliance, which is threatened by the risk of misinterpreting a requirement like PRT01, implementing PRT02 enhances the confidence in building compliant software. In this paper, assessing legal compliance is defined as the ability to provide justification for the choices made regarding certain implementation details. Since PRT02 is derived from a guideline issued by an authoritative entity (i.e., an authority with delegated power to adopt guidelines), it can be considered as a recommended best practice for achieving compliance. Thus, PRT02 facilitates the technical implementation of the right to portability according to the GDPR and provides a valid justification for compliance.

Inspired by the recent research initiatives on applying large language models (LLMs) in requirements elicitation [20] as well as in the legal domain [21], [22], [23], [24], this paper evaluates to what extent LLMs can assist requirements engineers in capturing detailed technical requirements related to data subject rights under the GDPR by leveraging additional sources.

Contributions. This paper makes the following contributions:

(1) A comprehensive list of GDPR requirements—manually extracted in close collaboration with legal experts—describing two key data subject rights, namely the *right of access* and the *right to portability* (hereafter referred to as ACC and PRT, respectively). As further elaborated in Section III, we chose these two rights because, despite being significantly relevant to software development and fundamental in the GDPR, they have not been discussed in-depth in the RE literature [17]. This implies potential shallow understanding and development that can have negative impact on compliance practices. The list contains 61 requirements pertinent to ACC and 47 requirements for PRT, extracted from the GDPR and other relevant sources, such as the guidelines issued by the European Data Protection Board (EDPB) [19]. We document these requirements and maintain explicit trace links to their respective references.

(2) To assess the effectiveness of LLMs in generating detailed requirements related to individuals' rights in the GDPR, we propose an automated approach, XTRAREG, standing for LLM-assisted Extraction of Regulatory Requirements. Through prompt engineering and retrieval augmented generation (RAG) technique [25], XTRAREG instructs a pre-trained LLM to generate detailed privacy requirements according to GDPR provisions and additional pre-defined legal sources.

(3) We empirically evaluate XTRAREG against the requirements extracted as part of contribution #1. Our evaluation shows that XTRAREG can automatically generate correct privacy requirements (primarily from the GDPR) with an accuracy of 81.8% for ACC and 85.7% for PRT. These requirements are further grounded in relevant legal references in 68.2% for ACC and 50% PRT, with additional fraction of incomplete references (27.3% for ACC and 21.4% PRT). These results demonstrate that LLMs can contribute to the requirements elicitation phase, which could in turn enhance the development practices of building compliant software. However, our experiments reveal the need for further research on guiding LLMs toward identifying details concerning the GDPR requirements.

Data availability. We make the list of manually extracted requirements and our evaluation material available in an online annex [26] and further release our tool on Zenodo [27].

Structure. The remainder of this paper is structured as follows: Section II provides background knowledge. Section III discusses the process conducted with the legal experts for extracting the privacy requirements. Section IV illustrates our LLM-assisted approach. Section V reports on our empirical evaluation. Section VI provides key insights derived from

our study. Section VII reviews the related literature. Finally, Section VIII concludes the paper.

II. Preliminaries

A. Legal background

EU law formally distinguishes between two different fundamental rights: the *right to privacy* of individuals and the *right to data protection* [28]. The former is primarily protected under Article 7 of the EU Charter of Fundamental Rights [5]. The latter is covered by Art. 16 of the Treaty on the Functioning of the EU, Art. 8 EUCFR, and, prior to the adoption of the GDPR, Directive 95/46/EC, known as the Data Protection Directive [29], which introduced *data protection principles* within EU law.

In 2016, the EU replaced the original Directive of 1995 with the General Data Protection Regulation (GDPR) [6], introducing a common and directly applicable legal framework[1] for all member states, with the aim of harmonizing their data protection principles and their application inside the EU Market. As highlighted in Section I, Chapter III of the GDPR states the rights of data subjects on their personal data. These rights include: 1) the right of access; 2) the right to data portability (hereafter shortened to "right to portability"); 3) the right to rectification; 4) the right to erasure; 5) the right to restrict the processing; 6) the right to object to the processing; and 7) the right not to be subject to decisions based solely on automated processing. In this paper, we focus on the first two of these data subject rights.

The *right of access* is a fundamental enabler of other rights related to personal data protection. As outlined in Art. 15 GDPR, this right allows data subjects to know whether their personal data are being processed and, if so, to access those data. It also grants several related rights, including the right to obtain information about recipients, processing purposes, duration and location of the processing, and the right to modify, erase, or lodge a complaint regarding their personal data. The *right to portability*, as outlined in Art. 20 of GDPR, implies the obligation for a data controller to recover the data subject's personal data in a machine-readable format and to expeditely transmit them to another data controller for reuse by the same data subject.

B. Technical background

NLP technologies have undergone significant and revolutionary changes in recent times. LLMs are currently dominating the state-of-the-art, outperforming traditional methods in several downstream tasks [30]. Language modeling essentially involves determining the probability of the next word in a given text sequence [31]. LLMs are massive pretrained models that have been exposed to vast amount of data, enabling them to obtain extensive knowledge about the world. Consequently, they can be leveraged without necessarily fine-tuning them on a specific task, making them a straightforward solution for addressing RE problems [32], [33], [34], [35]. Due to several limitations reported in the literature (e.g., hallucination), LLMs are best used as assistants to human experts.

Effectively instructing LLMs to solve a specific problem requires carefully designing prompts. In this work, we apply two different prompting techniques:

1) *Zero-shot learning* (`ZSL`) [36], [37] provides the LLM with explicit instructions on the task needed to be accomplished.
2) *Chain-of-Thought* (`CoT`) [38] guides the LLM to articulate reasoning behind each generated output.

We selected these techniques for the following reasons: (a) `ZSL` is the most straightforward option that a requirements engineer would use in practice; (b) `CoT` enforces transparency about the reasoning behind extracting the requirements, which is particularly relevant for our approach, as it is designed to assist human analysts. In this work, we do not use *few-shot learning*, another well-known prompting technique, due to the lack of training examples in our application domain.

We also leverage retrieval augmented generation (`RAG`) [25] to incorporate domain-specific knowledge during the prompting phase, effectively guiding the LLM toward the desired output. `RAG` builds on the traditional question-answering pipeline; to enhance the prompt, a question related to the original prompt is defined, and its answer provides the necessary context. The `RAG` vanilla pipeline consists of two modules, namely *indexing* and *retrieval and generation*. The indexing module prepares domain-specific documents in a suitable format for querying by loading the documents, splitting them into chunks, and storing their embeddings. The retrieval and generation module uses semantic similarity measures to retrieve the context (i.e., text sequence) most relevant to the predefined question. This context is then added to the original prompt to help the LLM generate more relevant results. We discuss the specific details relevant to our approach in Section IV.

III. Manual Requirements Extraction from GDPR

Data subject rights are clearly stated in the GDPR, albeit at a generic level. The exact implications and concrete implementations of these rights are specified in additional legal sources, such as the decisions of the European Court of Justice and the guidelines provided by bodies such as the European Data Protection Board (EDPB) [19]. All of the aforementioned sources contain requirements for systems dealing with personal data, whose identification is a fundamental step toward avoiding any violation of data subject rights, thus ensuring compliance to the GDPR.

This section outlines the process we conducted in close collaboration with subject-matter experts to extract data protection requirements derived from the GDPR and related legal sources.

As discussed in Section II, in this work we focus on legal requirements concerning two data subject's rights, namely the *right of access* and the *right to portability*. The reason behind this choice is that, despite the right of access being essential

[1]When the Directive 95/46 was in force, its concrete implementation would depend on each member state's transposition in its national legislative framework, with possible discrepancies among member states.

for enabling other data protection rights, it is not discussed in-depth in the RE literature [17]. The complexity of this seemingly straightforward right goes beyond what is stated in the GDPR, making appropriate requirements specification and implementation a challenge for software engineers. Regarding the right to portability, it is also relevant to software applications, e.g., when users switch to another online service or mobile app to obtain the same service. Furthermore, its implementation is expected to be reinforced with the recent EU Data Act [39]. Investigating the right to portability is valuable from a legal standpoint as most of the related literature focuses on its practical implications [40], but little attention has been given to its implementation.

Below, we describe the methodology followed for extracting data protection requirements, formulate some remarks on the process, and discuss the compliance scoring assigned to each requirement.

A. Methodology for Requirements Extraction

We extracted the data protection requirements on the right of access and the right to portability from various sources:

- *Mandatory legal sources*, such as legislation (e.g., GDPR) and judicial decisions (called *case law* in common law systems);
- *Advisory legal sources* (also known as *soft law*), such as guidelines issued by official bodies (e.g., the guidelines on the right to portability [19]);
- *Academic sources*, such as literature contributions (e.g., an academic paper describing the discipline on data portability [40]) and opinions of legal experts.

The requirements extraction process involved a total of six subject-matter experts (co-authors of this paper), including E1 and E2 who have substantial experience in RE and regulatory compliance, E3 who holds a PhD in legal informatics and more than five years of experience in interdisciplinary research topics combining law with software engineering, E4 who holds a PhD in criminal law and information technology law with a specific expertise on the application of the GDPR for data subjects' protection, E5 who holds a master degree in EU law, and E6 who holds a PhD in banking law and has field experience as compliance consultant and as certified Data Protection Officer in the financial domain.

To ensure the traceability of the extracted requirements, as a first activity E3, E4, and E5 identified and categorized the following legal sources to be used during the extraction process: *Applicable Legislation* (L) [6], *Official Guidelines* or *Technical Specifications* (G) [41], [19], [42], *Legal Literature* (T) [40], *Judicial Decisions* (J) [43], [44], and *Legal Expert* (E, corresponding to the opinion of E4/E5).

Following this, E4 extracted the legal requirements for the right of access, while E3 and E5 independently extracted the requirements for the right to portability and then consolidated their findings into a coherent set of requirements. The two sets of requirements were discussed, validated, and iteratively refined in four face-to-face sessions (with each session lasting at most two hours to avoid fatigue) involving primarily E3, E4, and E5, with occasional input from E1, E2, and E6. We note that we did not compute the inter-rater agreement since the extraction process was conducted mostly in online, interactive sessions. Disagreements, typically concerning refinements of requirements, were directly discussed and resolved. Finally, E1, E2, and E3 refined the list of requirements using a uniform template and consistent language, ensuring that no details from the legal sources were lost.

TABLE I: Excerpt of Privacy Requirements for Data Subjects' Rights Extracted from GDPR and Other Legal Sources

ID (Source)	Requirement
`ACC01` (L)	The system shall allow the data subject to request if data related to them are processed. [6, Art. 15.1]
`ACC03` (G)	The system shall allow the data subject to request access to the data undergoing processing related to them through a specific channel, appropriate and user-friendly, preferably within the own app or, failing that, through a link to an online feature. [42, § 53]
`ACC14` (GE)	The system shall only request full identification in exceptional cases. [41, p. 25]
`ACC36` (J)	The system shall reproduce extracts from documents or even entire documents or extracts from databases which contain, inter alia, the personal data undergoing processing, where the contextualisation of the data processed is necessary to ensure the data are intelligible. [44, §41]
`PRT01` (LG)	The system shall transfer to the data subject, upon a generic portability request, all personal data provided by the data subject. [6, Art. 20(1)][19, p. 4]
`PRT02` (GT)	The data transfer shall include all personal data that data controllers have collected upon consent or according to a contract, e.g., through GPS (location data), cookies, preferences, fitness data, purchasing history, credit card number, etc. [19, p. 10] [40, p. 199]

The extraction process resulted in 61 requirements for the right of access and 47 requirements for the right to portability. Table I shows a subset of the final requirements: requirements with prefix `ACC-` concern the right of access, whereas requirements with prefix `PRT-` concern the right to portability. We provide the complete list of requirements in an online annex [26]. The table shows, for each requirement, the legal source(s) from which the requirement was extracted, referring back to the categorization introduced above.

B. Observations

During the extraction of the requirements and the discussions for their refinement, the following observations emerged:

(O1) Best Practices. We observed *optional* requirements coming from sources beyond the GDPR. Although not compulsory by law, they depict ideal scenarios that would significantly reduce the risk of breaching the GDPR. For example, the guidelines on the right to portability [19] state that "it is a leading practice for all data controllers to implement tools to enable data subjects to select the relevant data they wish to receive and transmit and exclude, where relevant, data of other individuals". This statement is translated into requirement `PRT18` in our list (see annex [26]). Despite not being a binding obligation, complying with `PRT18` would reduce any risk of

breaching the GDPR by transmitting personal data concerning third parties. Implementing such a requirement in a software system largely depends on the confidence level that the data controller aims to achieve regarding the software's compliance. In our analysis, we label such requirements *best practices* to modulate compliance scoring, as we elaborate next.

(O2) Legal interpretation according to the application context. We observed that general-purpose regulations like the GDPR often express rather general and abstract norms which would translate into vague requirements. Translating these norms into actionable compliance requirements sometimes necessitates the exploration of other legal sources (such as national laws or sector-specific standards) for additional domain-specific knowledge related to the application context. To address this observation, we mark several requirements in our list as *vague*, indicating the need for further refinement. At this level of abstraction, these requirements are applicable in all EU member states for all domains under the scope of the GDPR. However, to specify implementable requirements, the exact legal details must be defined separately for each application context. For instance, requirement `PRT04` (see annex [26]) recommends the use of a "structured, commonly used and machine-readable format" to respond to a portability request, but it omits the format details, which may vary across application contexts and regions.

(O3) Alternative legal interpretations for the same requirement. We observed that the specification of a requirement can depend on the choice among different alternative interpretations. For instance, the definition of "data provided by the data subject" provided in `PRT02` in Table I contributes to determining the object of the transmission request in `PRT01`. For this requirement, we identified two different definitions clarifying this concept from two sources: the guidelines and the legal literature. From the perspective of compliance, both are considered correct, and selecting an interpretation would depend on the legal expert considering various criteria, including the application context. During the requirements extraction process we selected the interpretation which, among the alternatives, originated from the most authoritative source (i.e., official guidelines), while still maintaining the traceability to the other, less authoritative source (i.e., the literature).

C. Compliance Scoring of Legal Requirements

Our compliance scoring methodology is inspired by existing work on ranking transparency in the software documentation of EU online platforms [45]. As highlighted earlier, achieving compliance ultimately depends on the ability to robustly defend and argue the lawfulness of an implementation decision. Relying on established practices for building these arguments provides a stronger foundation than ad hoc decisions and opinions of individual legal experts.

For these reasons, to indicate the level of compliance confidence achieved by fulfilling a requirement, we assign a compliance score to each requirement extracted in our study, based primarily on their legal source. We identify two degrees of compliance: (i) *Minimal Compliance*, i.e., compliance to the minimal requirements of legislation (corresponding to *mandatory legal sources*: L, J), and (ii) *Recommended / Best Practices*, coming from *advisory and academic sources* (G, D, E), which ensure compliance with higher level of confidence. While an unfulfilled requirement on minimal compliance would surely lead to non-compliance with the law, thus raising a red flag (breach), an unfulfilled requirement on best practice would just raise a warning because arguing for compliance might be harder.

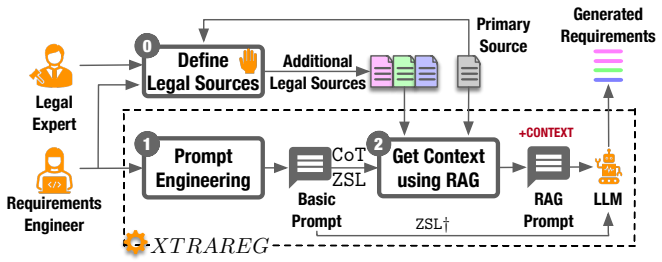

Fig. 2: Overview of XTRAREG

Hereafter, we refer to our manually extracted requirements as "reference requirements"

IV. LLM-assisted Approach for Extracting Detailed Privacy Requirements

Fig. 2. shows an overview of XTRAREG, our LLM-based approach designed to generate detailed privacy requirements. XTRAREG takes as input the primary legal source (GDPR in our case) and a pre-defined set of additional sources, and outputs a list of automatically generated detailed requirements. XTRAREG has a preparatory step (step 0 in Fig. 2), where legal experts and requirements engineers define the legal sources for extracting requirements. In our work, this step was performed manually with the legal experts, as explained in Section III. The core workflow of XTRAREG is then centered around two steps, elaborated below. In step 1, we define a set of prompts dedicated for generating detailed requirements. In step 2, we use these prompts to instruct the LLM to generate privacy requirements, given the input legal sources.

Step 1: Prompt Engineering. In the first step, we define a set of prompts for generating privacy requirements pertinent to the right of access and the right to portability under GDPR. The prompts are based on `ZSL` and `CoT` techniques. We show below the `ZSL` prompt to generate the requirements for the right of access[2]. These prompts aim to instruct the LLM to generate a list of requirements, with explicit mentions of the legal references and rationale for each generated requirement. Including this information in the output helps verify the accuracy of the LLM and its reasoning. For reusability, we include the remaining prompts in the annex [26].

Additionally, we designed a prompt (`ZSL†`), described below. Unlike the above prompts which emphasize GDPR as the main source for extracting requirements, the `ZSL†` prompt

[2]We use the word "elicit" in our instructions to the LLM where we actually mean "extract or generate". The rationale is to encourage the model to simulate the requirements elicitation phase and generate more detailed requirements.

explicitly instructs the LLM to refer to legal sources beyond GDPR. The rationale behind this prompt is to better understand whether the LLM can leverage its extensive knowledge for generating detailed requirements or whether it is strictly necessary to manually predefine the exact legal sources prior using LLMs (like we did in Step 0) in a similar context.

ZSL Prompt for ACC

```
Elicit detailed requirements on the right to
access from GDPR.
Write explicitly the text of each requirement.
Specify the exact reference for each
requirement.
Provide rationale for each requirement.
Give each requirement a unique ID using the
format "ACC" suffixed by a number.
Produce the output in JSON format.
```

ZSL† Prompt for ACC

```
Elicit comprehensive requirements related to
the right of access under the General Data
Protection Regulation (GDPR).
Instructions:
1. Identify and elaborate on the specific
requirements concerning the right of access
as stipulated in the GDPR.
2. Supplement these requirements with additional
authoritative sources beyond the GDPR, such as
guidelines from the European Data Protection
Board (EDPB) or relevant case law.
3. For each requirement, explicitly cite
the source, including therelevant article or
section from the GDPR, EDPB guidelines, or other
authoritativedocuments.
4. Provide a clear rationale for each
requirement, explaining its importance and
implications for data subjects and data
controllers.
```

Step 2: Get Context using RAG. In the next step, we integrate relevant contexts from the predefined legal sources into the prompts using a RAG vanilla pipeline (see Section II). Our RAG module is based on the default implementation provided by LangChain[3]. Specifically, it covers the following steps:

(1) Loading and indexing documents: We first use a pdf-loader to load and parse the textual content of a given legal source. We note that all additional legal sources in our study are provided in PDF format. Following the recommendations for handling text, we then split the document into chunks, each consisting of 1000 characters. We further maintain an overlap of 200 characters between chunks to preserve contextual coherence and avoid abrupt interruptions in the text. The resulting chunks are then transformed into embeddings and stored into a format that enables indexing and retrieving information. In our approach, we use the embeddings model provided by OpenAI. In particular, we use the 'text-embedding-ada-002' model, which generates embeddings of length 1536. The output of this step is an indexed representation of a given legal source, ready to be queried for relevant information.

(2) Retrieving relevant contexts: For each legal source, we prepare its corresponding indexed chunks as described above. To retrieve the relevant contexts, we then query the legal source using the following question, where *right* is replaced with *ACCESS* or *PORTABILITY* to cover the two rights. The resulting contexts are subsequently used to enhance our original prompts.

```
"What are the technical requirements for GDPR's
right to" + {right} + "?"
```

We develop the RAG module as a chain consisting of four elements:

- The first element is our customized question that is fed into a retriever to get the top-k relevant contexts (chunks). We used the default value for k as provided by LangChain, which is 4. The similarity metric applied in the retrieval step is based on Euclidean distance, i.e., the lower the value, the more similar a context is to a specific query.
- The second element is the prompt including our instructions combined with the top-k contexts and the question. For each right, we use the ZSL and CoT prompts resulted from Step 1. We note that the ZSL† prompt is not passed through the RAG module, as we discuss in Section V.
- The third element is the LLM; as discussed in Section V, in our experiments we used two models, namely GPT3.5 and GPT4o. We set the temperature to 0.3. We chose this value based on our preliminary experiments where we observed that allowing some variability in the LLM answers leads to more elaborated results, compared to setting temperature to 0 from the beginning which is more likely to lead to more errors, as acknowledged in the literature [46].
- Finally, the last element is a parser that transforms the raw output of the LLM into plain text.

In summary, using RAG ensures that the LLM will generate according to our instructions a list of requirements derived from the GDPR and informed by the other legal sources. To obtain a broader set of requirements, we apply the RAG module on each legal source, defined in Step 0. The final output of our approach represents the concatenated list of requirements generated across the multiple legal sources.

Implementation. We implemented XTRAREG using Python 3.9.12 and Jupyter Notebooks. Specifically, we prompted the GPT models using ZSL and CoT via the OpenAI API (1.14.0). We implemented the RAG pipeline using the LangChain libraries: *PyPDFLoader* available in *community.document_loaders* (0.2.10) for loading the legal sources, *text_splitters* (0.2.2) for splitting them into chunks, *Chroma* available in *community.vectorstores* (0.2.10) for creating the indexed representation, *core* (0.2.26) for implementing the parsing functionalities, *openai* (0.1.20) integrates the OpenAI GPT models into the RAG chain. The implementation details can be found in our replication package [26].

V. Evaluation

In this section, we report on our empirical evaluation, through which we investigated two research questions (RQs):

[3]https://python.langchain.com/docs/tutorials/rag/

RQ1. Which configuration of XTRAREG yields the best accuracy in extracting privacy requirements for GDPR rights of access and to portability? Extracting a complete set of privacy requirements is paramount for ensuring that software systems comply with regulations. XTRAREG utilizes LLMs to automatically generate requirements from GDPR and other legal sources; various technologies can be integrated with LLMs to enhance this extraction process. RQ1 assesses the accuracy of XTRAREG in the requirements extraction task, when using different configurations, each consisting of an LLM and a prompting technique.

RQ2. What is the impact of using `RAG` on the performance of XTRAREG? LLMs are trained on extensive data, likely including the predefined legal sources used by experts (see Section III). In RQ2, we investigate the impact of explicitly exposing the LLM to these specific legal sources using the `RAG` technique, compared to explicitly instructing the LLM to leverage its extensive knowledge as needed to generate a more comprehensive set of requirements.

A. Accuracy of Privacy Requirements Extraction (RQ1)

To answer RQ1, we assessed various configurations of prompting techniques and LLMs to identify the most effective combination.

1) LLM settings: We assessed the accuracy of XTRAREG in retrieving detailed privacy requirements using two alternative prompting techniques, namely `ZSL` and `CoT` (see Section II) and two models, namely GPT3.5 (`gpt-3.5-turbo-0125`) and GPT4o (`gpt-4o-2024-08-06`) a widely-used family of LLMs provided by OpenAI (https://platform.openai.com/docs/models). Our motivation behind selecting those GPT models is driven by their reasonable performance (e.g., [47]) as well as their popularity due to being easily accessible through the web interface. Furthermore, testing two different model generations enables us to better analyze how the evolution of LLMs impacts performance in the legal domain.

2) Accuracy metrics: Evaluating automatically generated text effectively remains an open research topic in the NLP literature [48], [49]. Recent literature uses LLMs to check themselves [50] or other LLMs' outputs [51]. Such evaluation methods are not suitable in our context for two reasons. First, LLMs are prone to self-enhancement bias (i.e., they tend to favor their own output [51]) which might lead to false conclusions about their performance. Second, unlike conventional tasks like text classification or information extraction where output can be clearly categorized as correct or incorrect, evaluating automatically generated legal requirements requires legal expertise to better decide whether these generated requirements convey equivalent interpretation from a legal standpoint (beyond the semantic relatedness). We therefore use a dual assessment method involving automated accuracy metrics and human evaluation, as explained next. We first assessed the generated output using fully automated accuracy metrics that are popular in both NLP and SE research [52], [53], [54], [55]. Specifically, we used three n-gram-based metrics, namely *BLEU* [56], *ROUGE-L* [57], and *METEOR* [58], and one semantic similarity metric, namely *BERTScore* [59]. BLEU is a precision-oriented metric, which assesses how many n-grams in the generated text t also appear in a reference text r. ROUGE is a recall-oriented metric, which assesses how many n-grams of t have been captured in r. METEOR considers synonyms besides original text and computes a score based on the harmonic mean of precision and recall. *BERTScore* assesses the semantic relatedness between t and r.

Then, we conducted an extrinsic assessment with a single legal expert who has expertise in both law and engineering necessary for this evaluation. Inspired by recent work on evaluating LLMs in the legal domain [60], [22], [61], the expert assessed each generated requirement, the associated proposed reference, and the provided rationale using the following criteria:

Correctness: A requirement is *correct* if it is factually correct and relevant. It is *partially correct* if it is incomplete or not fully capturing the requirement. Otherwise, it is *incorrect*. For a more fine-grained assessment, the expert assigned a numerical value (1–5) to each requirement according to the following criteria: The value *5* is assigned when the requirement is correctly formulated; *4* is assigned when the requirement is provided from the wrong perspective (i.e., that of the data subject and not that of the system); *3* is assigned when the requirement contains a generic reference to a legal act such as a contract or agreement, or if parts of the original norm are missing or left implicit; *2* is assigned when the requirement contains a generic reference to the rights of individuals; *1* is assigned when the requirement is *incorrect*, e.g., because its meaning depends on a reference to a normative source (i.e., an unresolved legal reference), or it indicates the wrong action. Values *5* and *4* indicate *correct*, while values *3* and *2* indicate *partially correct*.

Groundedness: A reference is *grounded* if it is valid and relevant to legal documents. It is *misgrounded* if it does not support the requirement, *ungrounded* if no reference is provided, or *incomplete* if not fully referenced.

Plausibility: A rationale is *plausible* if it is reasonable and supports the requirement; otherwise, it is *implausible*.

We opted for a coarse-grained assessment in the latter two due to the binary nature of the defined labels.

The expert further labeled each generated requirement with the equivalent reference requirement(s), according to which we then computed **Coverage**, indicating how many requirements the LLM could identify from our reference requirements.

3) Methodology: Our automated evaluation aims to examine side-by-side the different configurations of XTRAREG. To obtain the results, we executed each configuration (e.g., GPT4o + `ZSL`) by injecting the most relevant context from applicable legal sources into the prompt. To account for output variations of LLMs, we run each configuration twice. Furthermore, we run XTRAREG separately for each right as not all legal sources defined in this work are relevant for both rights.

For each configuration, we evaluated the generated requirements by comparing them against the reference require-

ments (see Section III). For a given metric (e.g., BLEU), we computed the value by comparing a generated requirement t against all reference requirements in our ground truth, denoted by $\mathcal{H}$ (e.g., $\text{BLEU}_1(t, r_1)$, $\text{BLEU}_2(t, r_2)$, ...), and assigned t the resulting maximum score. We then computed the average score for each configuration across all generated requirements. While the maximum score implicitly assumes a one-to-one relation between t and r (which might not be correct as t might capture multiple reference requirements), it provides a fair assessment across the different configurations using their best possible accuracy values.

4) Results (Automated Metrics): The top half of Table II shows the scores (with RAG) and the number of unique requirements (N) generated by each configuration for the rights of access (ACC) and portability (PRT). The table shows the average values over two runs. We focus our analysis on GPT4o since, as explained earlier, GPT3.5 failed to generate adequate number of requirements to provide meaningful assistance.

The scores for ACC are consistently higher than those for PRT across various configurations. This difference can be attributed to the extensive availability of legal sources for ACC, whereas PRT has not yet gained widespread recognition. This observation is further supported by the fact that GPT4o outperforms GPT3.5 across all metrics for ACC, while performing similarly for PRT. This suggests that the training data for both models, even the newer one, lacks adequate information on PRT. The advantage of GPT4o over GPT3.5 is however its ability to generate remarkably more requirements (e.g., 3 vs 22 for ACC when using GPT4o + CoT), indicating a greater tendency toward completeness. We therefore focus the rest of our analysis on GPT4o.

GPT4o achieves relatively lower BLEU scores compared to ROUGE scores, suggesting that the generated requirements, compared with the reference requirements, are less verbatim (i.e., fewer words in the generated requirements also appear in the reference requirements). The generated requirements are also shorter since ROUGE measures the ratio of overlapping n-grams with respect to the n-grams in the generated requirements. Higher METEOR scores compared to BLEU, on the other hand, emphasize that the generated requirements contain synonyms (instead of verbatim words) from the reference requirements to convey similar meanings, as evidenced by the high BERTScores.

Besides the notably higher number of generated requirements, our evaluation shows that GPT-4o performs consistently well across various prompting techniques. However, differences may emerge in the model reasoning, particularly in the rationale provided for each requirement, typical for CoT. To better understand the model behavior, we manually analyzed the generated outputs of GPT4o for both prompting techniques, namely ZSL and CoT.

5) Results (Extrinsic Assessment): Using the criteria defined earlier, we qualitatively evaluated the generated requirements, the cited legal references, and the rationale. The legal expert analyzed a total of 42 requirements for ACC and 32 requirements for PRT, generated by GPT4o[4]. The top half of Table III shows the results of the qualitative analysis for this RQ. Specifically, for each configuration, the table lists the total number of generated requirements (N), the number of duplicate requirements (D) out of N, the *coverage* (V) with respect to the reference requirements, the total number of requirements labeled according to the above-defined criteria, namely *correct* (C1), *partially correct* (C2), *incorrect* (C3), *grounded* (G1), *incomplete* (G2), *misgrounded* (G3), and *ungrounded* (G4). We omit *plausibility* since the expert found all provided rationale to be consistently *plausible* for both rights.

The results show that XTRAREG maintained a consistent *coverage* (V) scores for ZSL and CoT. We further observed that the reference requirements identified by XTRAREG are identical for both prompting techniques. This suggests that the prompting technique has little impact in our context. Therefore below we discuss the results of ZSL only.

XTRAREG generated 22 requirements for ACC, of which 36.4% (8/22) are duplicates, and 72.7% (16/22) correspond to reference requirements in our manually extracted list (i.e., coverage = 16/61). We note that one generated requirement can be mapped to multiple reference requirements. Our analysis further reveals that all identified requirements are exclusively from the GDPR (particularly, Art. 15: *"Right of access by the data subject"*). Our reference requirements contained a total of 33 requirements from the GDPR, 18 of which are from Art. 15 while the rest originates from other articles, broadly related to data subject rights, e.g., Art. 12 on *"Transparent information, communication and modalities for the exercise of the rights of the data subject"*. For PRT, XTRAREG generated 14 requirements, of which 28.9% (4/14) are redundant, and 57.1% (8/14) are relevant to our reference requirements. In line with our observation above, XTRAREG generated requirements from the GDPR, Art. 20 which is explicitly about the "Right to data portability". We note that 18 requirements (out of 47) originate from the GDPR (Art. 20 and other articles). The above analysis suggests that LLMs tend to distill knowledge that could strictly be relevant to the right, but they fail to generate requirements for ACC or PRT as being part of the general data subject rights, hence neglecting the other articles in the GDPR (e.g., Art. 12 mentioned above).

Table III further shows that 18 out of the 22 generated requirements by XTRAREG for ACC were labeled as correct (C1) or partially correct (C2) by the expert (corresponding to 81.8%); in the case of PRT, we have 12 out of the 14 requirements (corresponding to 85.7%). As shown in the table, XTRAREG often suggests correctly grounded references (G1): 68.2% (15/22) for ACC and 50.0% (7/14) for PRT. However, a fraction of 27.3% (6/22) of the proposed references are deemed incomplete (G2) for ACC (for PRT, 21.4% (3/14)). Incomplete references are primarily due to not proposing the corresponding GDPR articles, indicating the limited capability of LLMs in pinpointing what is normative in the legisla-

[4]The expert also analyzed 14 requirements generated by GPT3.5, but, due space limitations, we only report them in the online annex [26].

TABLE II: Accuracy of Requirements Extraction (RQ1) and Ablation Study Results (RQ2)

Configuration	ACC					PRT				
	N	BLEU	ROUGE	METEOR	BERTScore	N	BLEU	ROUGE	METEOR	BERTScore
GPT3.5 + ZSL (w)	3±0.0	0.165±0.00	0.461±0.00	0.489±0.00	0.863±0.00	3±0.0	**0.102**±0.00	0.357±0.00	**0.419**±0.00	0.837±0.00
GPT3.5 + CoT (w)	3±0.0	0.165±0.00	0.461±0.00	0.489±0.00	0.863±0.00	3±0.0	**0.102**±0.00	0.357±0.00	**0.419**±0.00	0.837±0.00
GPT4o + ZSL (w)	**22.5**±0.7	0.242±0.01	0.494±0.01	**0.614**±0.01	**0.881**±0.00	11.5±2.1	0.073±0.00	**0.382**±0.01	0.401±0.01	**0.848**±0.00
GPT4o + CoT (w)	22±2.8	**0.255**±0.01	**0.504**±0.00	0.612±0.00	0.880±0.00	**14.5**±2.1	0.062±0.01	0.358±0.01	0.381±0.01	0.837±0.01
GPT3.5 + ZSL (w/o)	4.7±0.8	0.069±0.05	0.372±0.05	0.370±0.10	0.838±0.02	4.2±0.6	0.095±0.03	**0.382**±0.03	0.383±0.05	0.837±0.01
GPT3.5 + ZSL† (w/o)	5.9 ±1.9	0.108±0.04	0.376±0.03	0.442±0.03	0.853±0.01	4.8±0.4	0.090±0.03	0.365±0.03	0.391±0.03	0.842±0.01
GPT3.5 + CoT (w/o)	4.4±0.8	0.082±0.05	0.377±0.06	0.385±0.12	0.839±0.02	3.8±0.6	0.097±0.03	0.370±0.03	0.388±0.05	0.837±0.01
GPT4o + ZSL (w/o)	8.7±1.9	0.137±0.10	0.403±0.09	0.410±0.19	0.848±0.03	7±1.5	0.056±0.04	0.330±0.05	0.283±0.10	0.819±0.03
GPT4o + ZSL† (w/o)	8.5 ±2.7	0.142±0.06	0.375±0.04	0.464±0.11	0.847±0.02	8.9±1.5	0.072±0.04	0.346±0.05	0.325±0.08	0.833±0.02
GPT4o + CoT (w/o)	10.6±2.1	0.115±0.11	0.382±0.10	0.361±0.21	0.836±0.04	6.8±0.9	0.070±0.04	0.345±0.05	0.300±0.09	0.824±0.02

N: Number of unique requirements, w: with RAG, w/o: ablation study without RAG.

tion [62]. A comparable proportion of the suggested references for PRT (28.6%, 4/14) are deemed misgrounded (G3). This confirms our previous observation about PRT being not yet popular. Despite the limitations of LLMs in the legal domain, as we elaborate in Section VI, the results reported above show a promising research trajectory which requires further investigation primarily for improving the coverage.

6) Answer to RQ1: XTRAREG achieves the best accuracy when using the GPT4o model regardless of the applied prompting technique; we selected ZSL for simplicity. With this configuration, XTRAREG generates privacy requirements strictly from the GDPR: 81.8% of the requirements for ACC are correct an 68.2% are grounded in valid legal references, whereas 85.7% of the generated requirements for PRT are correct and 50% are grounded. XTRAREG can suggest incomplete legal references (27.3% for ACC and 21.4% PRT). Despite the LLMs limitations in the legal domain with respect to providing comprehensive requirements, we believe these results can serve as a foundation for future research.

B. Evaluating the Impact of RAG (RQ2)

To answer RQ2, we conducted an ablation study where we applied XTRAREG without utilizing the RAG technique.

1) LLM settings: Here, we let the LLM generate, based on our prompts, a set of requirements by mining the massive knowledge it was trained on. As in RQ1, we evaluated the alternative prompts ZSL and CoT. We further introduced ZSL†, a variant of ZSL where we explicitly instruct the LLM to use additional legal sources. The rationale behind experimenting with this prompting technique is three-fold. First, ZSL is an intuitive technique that can be used by a requirements engineer without having to define which legal sources are relevant to their application context. Second, the massive knowledge of LLM would likely contain the legal sources we defined in our qualitative study (in Section III), yet the question is whether the LLM would be able to pinpoint such sources. Finally, driven by the results in RQ1, we aim to assess whether leveraging a broader set of legal references through ZSL† would improve the comprehensiveness of the generated requirements.

2) Accuracy metrics: We used the same metrics as in RQ1.

3) Methodology: To enable a meaningful comparison with our analysis in RQ1, we used the same prompts, but without the RAG context. To better understand the model behavior, we run each configuration ten times encouraged by the lower cost of the experiments in RQ2 compared to those in RQ1 (exacerbated by running the RAG pipeline on several legal documents). We then computed the metrics for each configuration following the same procedure described in RQ1.

4) Results (Automated Metrics): The bottom half of Table II (lines marked with "w/o") shows the results of the ablation study, listing the mean and standard deviation for each configuration.

When RAG is not used, XTRAREG exhibits a drop in its performance across the metrics for both ACC and PRT, with one exception where GPT3.5 produced comparable ROUGE values for PRT. This observation proves the usefulness of exposing legal sources to generate requirements that are more closely aligned with our ground truth. The LLM's behavior remained, however, consistent with our analysis in RQ1, e.g., the BLEU score is relatively low compared with ROUGE or METEOR for the same reasons discussed earlier. Another notable observation is that GPT3.5 performs relatively better than GPT4o across all metrics for PRT in the ablation study. The table further shows that, unlike GPT4o, GPT3.5 generated, on average, more requirements with RAG than without.

5) Results (Extrinsic Assessment): To understand the reliability of the automated evaluation, we analyzed in our extrinsic assessment the output files on both extremes, i.e., those with the best (and worst) scores. To facilitate comparisons, in particular between exposing specific legal sources (i.e., with RAG) vs letting the LLM freely leverage its massive knowledge (i.e., without RAG), in Table III (bottom half) we only show a subset of the results generated by GPT4o using ZSL and ZSL†; the complete results are included in the online annex [26].

Without RAG, XTRAREG generated 10 requirements for ACC when using ZSL, and 12 when explicitly instructed through ZSL† to leverage external resources beyond GDPR. These generated requirements are distinct and correspond to 11 and 12 reference requirements, respectively. It is worth noting that all requirements generated using ZSL are from Art. 15, whereas the use of ZSL† triggered generating requirements

TABLE III: Results of Qualitative Analysis with RAG (RQ1) and without RAG (RQ2).

Configuration	**ACC** (61 Reference Requirements)										**PRT** (47 Reference Requirements)									
	N	D	V	C1	C2	C3	G1	G2	G3	G4	N	D	V	C1	C2	C3	G1	G2	G3	G4
GPT4o + ZSL (w)	22	8	16	14	4	4	15	6	1	0	14	4	8	5	7	2	7	3	4	0
GPT4o + ZSL (w/o)	10	0	11	0	9	1	0	0	10	0	6	0	2	0	6	0	3	1	2	0
GPT4o + ZSL† (w/o)	12	0	12	2	4	6	8	2	2	0	11	0	6	7	3	1	11	0	0	0

N: Number of generated requirements, D: duplicates with respect to N, V: Coverage with respect to r, Correctness (C1: Correct, C2: Partially correct, C3: Incorrect), Groundedness (G1: Grounded, G2: Incomplete, G3: Misgrounded, G4: Ungrounded). w: with RAG, w/o: ablation study without RAG.

from other articles of the GDPR. For PRT, XTRAREG generated six requirements using ZSL and 11 using using ZSL† (no duplicates in either case). With ZSL, only two reference requirements were identified, both from Art. 20 in the GDPR. With ZSL, however, six reference requirements were identified from various articles in the GDPR as well as the EDPB guidelines (see § III). While RAG seems to be essential for better *coverage*, compared to ZSL†, it falls short on guiding the LLM toward generating requirements beyond the GDPR.

As can be seen in the table, nine out of the 10 requirements generated for ACC using ZSL are deemed partially correct, yet all their associated references are misgrounded. Using ZSL†, on the other hand, resulted in only six out of 12 correct or partially correct requirements. However, while requirements might not be correct, the majority of them are well grounded in the legal sources. This shows that when the LLM consults undefined legal sources, it may suggest requirements that are not directly aligned with the manually extracted ones. Similar observations can be made on the generated requirements for PRT, though using ZSL† resulted in more correct (or partially correct) requirements than when ACC was used. Additionally, using ZSL (rather than ACC) yielded better *groundedness*.

6) Answer to RQ2: Despite the redundant requirements, RAG significantly impacts XTRAREG's performance. However, RAG alone is not adequate for generating a comprehensive list of privacy requirements beyond the GDPR. Nonetheless, we believe our observations are beneficial foundation for advancing the research in this direction.

C. Threats to Validity

Internal Validity. The main threat to internal validity is subjectivity both in interpreting the legal sources as well as qualitatively evaluating the results. To mitigate this threat, we collaborated with subject-matter experts. Our manual extraction involved several experts with legal background; the extrinsic evaluation was collectively done by two experts. All experts involved in the qualitative analyses had little to no exposure on the implementation details. We further make our results publicly available and open for scrutiny.

Another consideration is the output variability which is a known issue when developing LLM-based solutions. In particular, it is opaque when and how GPT models are updated and how these updates can affect their behavior [63]. To reduce the effect of this threat, we collected our results from two different generations of GPT models. We further run our experiments over multiple iterations in two distant time frames. That said, further experimentation could be beneficial to improve the generalizability of our findings.

Construct Validity. Automatically generated requirements cannot be mapped one-to-one to the manually extracted ones. To ensure the validity of our findings, we defined, based on existing literature, precise and explicit evaluation criteria and performed human evaluation on a broad set of the results.

VI. Discussion

What to pay attention to when utilizing LLMs in the legal domain? Our qualitative analysis of the automatically generated requirements has revealed, from a legal perspective, limitations of LLMs that must be accounted for in this context. Specifically, we make the following generic observations regarding the correctness of the requirements:

(i) The LLM often adopts an incorrect perspective when formulating requirements, e.g., instead of "the system shall", the requirements may use phrases such as "the exercise of the right [...]".
(ii) The LLM frequently generates generic requirements that (1) are not detailed beyond the legal formulation, and (2) sometimes overlook exceptions.
(iii) Allowing the LLM to more freely mine its legal knowledge (instead of explicit exposure of pre-defined legal references) results in more compact—and occasionally clearer—requirements. However, these requirements are generally less contextualized and more prone to containing incorrect actions.
(iv) In some cases, the LLM would produce non-actionable requirements, providing generic descriptions or verbatim excerpts of the legal text.

Regarding references, the LLM (possibly influenced by practices in common law systems, as captured in its training data) often cites recitals exclusively, leading to incomplete output when applied in civil law systems such as the EU.

Some of the above limitations, such as the wrong perspective, can be addressed with more advanced prompting strategies, whereas others, like correcting the actions, would require involving legal experts (possibly with automated support in place) in the process of extracting legal requirements.

How XTRAREG contributes to practical implementation of GDPR individual's rights in software. Our work on extracting legal requirements for the right of access and the right to portability also highlight the complexity and the wide range of implications that those rights (and data subject rights in general) can have. Such implications go beyond transparency [64]

and mandatory disclosures in privacy policies [13] as per GDPR Articles 12 and 13 [6]. They also imply concrete implementation considerations enabling users to exercise their rights, ideally directly through dedicated support that need to be provided to them, or indirectly through request forms, followed by transparent and diligent back-end processing by companies, which need to be supported to.

Our work provides the research community with a comprehensive list of detailed requirements pertinent to two key rights in the GDPR. Our list can be used by requirements engineers to better implement the rights of access and to data portability. Since our proposed approach XTRAREG performs well by just relying on the LLM's legal knowledge (i.e., without RAG), it can be used by engineers to generate more detailed legal requirements when the accessibility to legal experts is scarce. Legal experts can also use XTRAREG to enhance efficiency in navigating through legal sources. As noted earlier, the LLM can provide verbatim excerpts from the legal documents. While not useful as-is, these excerpts would often pinpoint the location of relevant information in the document. That said, further research is needed to investigate the other rights and better align the software development in practice with the legal perspective on compliance.

How to generalize XTRAREG beyond the regulatory framework used in this paper? Though XTRAREG was originally designed and evaluated for specific rights under the GDPR, it can be applied, beyond this context, for extracting requirements for other compliance-related goals derived either from the GDPR or from other primary legal sources. Reusability entails adapting the following parameters: (1) The input primary source and subsequently the additional relevant legal references (step 0 must be repeated); (2) the prompts, e.g., by replacing the mentions of the specific rights with the desired legal concepts.

VII. Related Work

Below, we review the RE literature on requirements elicitation, the most relevant to our work on requirements generation.

Requirements elicitation has been extensively studied in RE [65], [66]. In the context of regulatory compliance, the extraction of compliance-related information from legal texts (an activity closely related to requirements elicitation) has been widely investigated. Existing work advocates for collaboration with legal experts [7], [12]. Both manual and automated methods based on traditional NLP/ML techniques have been proposed for analyzing privacy requirements from the Health Insurance Portability and Accountability Act (HIPAA) [67], [68]. Similar techniques have been applied to, e.g., other legislation [9], [10] and privacy policies [64], [13]. Our work differs in two key aspects. First, we collaborate closely with legal experts to extract a comprehensive list of detailed requirements for the GDPR rights of access and portability, incorporating legal sources beyond the GDPR. Second, we explore the feasibility of automated assistance using recent NLP technologies.

The RE literature has shifted toward using LLMs for addressing various challenges, emphasizing the different phases of requirements engineering [32], more particularly requirements elicitation [69], [70], [71], [72]. Existing work has also investigated the use of LLMs for privacy and data protection compliance, e.g., [73], [74]. Ioannidis et al. [75] investigated the effectiveness of LLMs in generating obligations from legislation. They used GPT-4 to create a list of obligations from various legislative and regulatory material. Compared to their work, we utilize RAG to explicitly expose the LLM to a specific set of regulations. We further challenge the LLM with a well-defined use case (individuals' rights in the GDPR) to better assess its performance in assisting human analysts.

The work closest to ours is by Ronanki et al. [76], which explored the potential of ChatGPT for assisting in general requirements elicitation. They prompted ChatGPT to generate requirements targeting specific quality attributes of trustworthy AI, including accuracy, robustness, and privacy. These generated requirements were then evaluated against manually defined requirements. Our work presents a complementary, yet distinct contribution. The privacy requirements generated by Ronanki et al. concern privacy in general; these requirements can be useful, but their refinement requires multiple iterations. In contrast, in this work, by focusing exclusively on privacy requirements, we were able to conduct an in-depth analysis of various legal sources and derive detailed technical requirements. Additionally, we investigated the necessity of integrating domain-specific knowledge into LLMs using RAG technology within the legal domain.

VIII. Conclusion

In this paper we have extracted, in close collaboration with legal experts, a list of requirements pertinent to the rights of access and portability in the General Data Protection Regulation (GDPR). To obtain detailed requirements, our extraction process incorporated relevant legal sources beyond the GDPR. Additionally, we have proposed XTRAREG, an automated approach that utilizes LLMs to support the requirements extraction activity. Our empirical evaluation shows that, despite the low coverage compared to our manually defined requirements, XTRAREG can generate correct ACC requirements with an accuracy of 81% and PRT requirements with an accuracy of 85.7%. XTRAREG further cites relevant legal references with an accuracy of 68.2% for ACC and 50% for PRT. We believe this work contributes to increasing awareness on the concrete implementation of data subject rights, which is an essential step toward better compliance against privacy regulations such as GDPR.

In the future, we plan to conduct user studies with engineers to assess the practical usefulness of XTRAREG. We would like to also investigate other concepts under GDPR to advance this research direction. Finally, we plan to investigate the impact of the temperature parameter on the LLMs capabilities in the legal domain.

Acknowledgment

This research was funded in whole, or in part, by the Luxembourg National Research Fund (FNR), under grants numbers NCER22/IS/16570468/NCER-FT and C23/IS/17958091/PLAITO.

References

[1] E. Felten, M. Raj, and R. Seamans, "Generative AI requires broad labor policy considerations," *Commun. ACM*, vol. 67, no. 8, pp. 29–32, 2024.

[2] R. Bhayana, "Chatbots and large language models in radiology: a practical primer for clinical and research applications," *Radiology*, vol. 310, no. 1, p. e232756, 2024.

[3] W. Rong and Z. Yu, "Do AI chatbots improve students learning outcomes? evidence from a meta-analysis," *Br. J. Educ. Technol.*, vol. 55, no. 1, pp. 10–33, 2024.

[4] W. Seymour, X. Zhan, M. Coté, and J. M. Such, "A systematic review of ethical concerns with voice assistants," in *Proceedings of AIES 2023*. ACM, 2023, pp. 131–145.

[5] European Union, *Charter of Fundamental Rights of the European Union*. Brussels: European Union, 2010, vol. 53.

[6] The European Parliament and the Council of the European Union, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," 05 2016.

[7] T. D. Breaux, M. W. Vail, and A. I. Antón, "Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations," in *Proceedings of RE 2006*. IEEE, 2006, pp. 46–55.

[8] S. Ghanavati, D. Amyot, and A. Rifaut, "Legal goal-oriented requirement language (legal GRL) for modeling regulations," in *Proceedings of MiSE 2014*. ACM, 2014, pp. 1–6.

[9] N. Zeni, N. Kiyavitskaya, L. Mich, J. R. Cordy, and J. Mylopoulos, "Gaiust: supporting the extraction of rights and obligations for regulatory compliance," *Requir. Eng.*, vol. 20, no. 1, pp. 1–22, 2015.

[10] A. Sleimi, N. Sannier, M. Sabetzadeh, L. C. Briand, M. Ceci, and J. Dann, "An automated framework for the extraction of semantic legal metadata from legal texts," *Empir. Softw. Eng.*, vol. 26, no. 3, p. 43, 2021.

[11] J. Tom, E. Sing, and R. Matulevicius, "Conceptual representation of the GDPR: model and application directions," in *Proceedings of BIR 2018*, ser. Lecture Notes in Business Information Processing, vol. 330. Springer, 2018, pp. 18–28.

[12] D. Torre, M. Alférez, G. Soltana, M. Sabetzadeh, and L. C. Briand, "Modeling data protection and privacy: application and experience with GDPR," *Softw. Syst. Model.*, vol. 20, no. 6, pp. 2071–2087, 2021.

[13] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. C. Briand, "AI-enabled automation for completeness checking of privacy policies," *IEEE Trans. Software Eng.*, vol. 48, no. 11, pp. 4647–4674, 2022.

[14] M. I. Azeem and S. Abualhaija, "A multi-solution study on GDPR AI-enabled completeness checking of dpas," *Empir. Softw. Eng.*, vol. 29, no. 4, p. 96, 2024.

[15] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An empirical evaluation of GDPR compliance violations in android mhealth apps," in *Proceedings of ISSRE 2020*. IEEE, 2020, pp. 253–264.

[16] M. Hatamian, S. Wairimu, N. Momen, and L. Fritsch, "A privacy and security analysis of early-deployed COVID-19 contact tracing android apps," *Empir. Softw. Eng.*, vol. 26, no. 3, p. 36, 2021.

[17] C. Negri-Ribalta and M. L.-P. C. Salinesi, "Understanding the GDPR from a requirements engineering perspective — a systematic mapping study on regulatory data protection requirements," *Requir. Eng.*, pp. 1–27, 2024.

[18] noyb - European Center for Digital Rights, "GDPR: a culture of non-compliance? Numbers of evidence-based enforcement efforts," Accessed Sep. 24, 2024 [Online], 2024. [Online]. Available: https://noyb.eu/sites/default/files/2024-01/GDPR_a%20culture%20of%20non-compliance_2.pdf

[19] Article 29 Data Protection Working Party, "Guidelines on the right to data portability," 2017. [Online]. Available: https://ec.europa.eu/newsroom/article29/items/611233

[20] B. Görer and F. B. Aydemir, "Generating requirements elicitation interview scripts with large language models," in *Proceedings of RE 2023 - Workshops*. IEEE, 2023, pp. 44–51.

[21] D. Bernsohn, G. Semo, Y. Vazana, G. Hayat, B. Hagag, J. Niklaus, R. Saha, and K. Truskovskyi, "Legallens: Leveraging llms for legal violation identification in unstructured text," in *Proceedings of EACL 2024*. ACL, 2024, pp. 2129–2145.

[22] F. Contini, "Unboxing generative AI for the legal professions: Functions, impacts and governance," *International Journal for Court Administration*, vol. 15, no. 2, 2024.

[23] A. S. Kwak, C. Jeong, G. Forte, D. E. Bambauer, C. T. Morrison, and M. Surdeanu, "Information extraction from legal wills: How well does GPT-4 do?" in *Proceedings of EMNLP 2023*. ACL, 2023, pp. 4336–4353.

[24] F. Yu, L. Quartey, and F. Schilder, "Exploring the effectiveness of prompt engineering for legal reasoning tasks," in *Proceedings of ACL 2023*. ACL, 2023, pp. 13 582–13 596.

[25] P. S. H. Lewis and others, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of NeurIPS 2020*, 2020.

[26] S. Abualhaija, M. Ceci, N. Sannier, D. Bianculli, S. Lannier, M. Siclari, O. Voordeckers, and S. Tosza, "Online Annex to the paper: "LLM-assisted Elicitation of Regulatory Requirements: A Case Study on the GDPR"," https://doi.org/10.6084/m9.figshare.27187557, 2025.

[27] ——, "Artifact associated with "LLM-assisted extraction of regulatory requirements: A case study on the GDPR"," https://doi.org/10.5281/zenodo.15668459, 2025.

[28] R. Gellert and S. Gutwirth, "The legal construction of privacy and data protection," *Comput. Law Secur. Rev.*, vol. 29, no. 5, pp. 522–530, 2013.

[29] The European Parliament and the Council of the European Union, "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," 10 1995.

[30] P. A. Chitale, J. P. Gala, and R. Dabre, "An empirical study of in-context learning in llms for machine translation," in *Proceedingns of ACL 2024*. ACL, 2024, pp. 7384–7406.

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Prentice Hall, 2020.

[32] C. Arora, J. Grundy, and M. Abdelrazek, *Advancing Requirements Engineering Through Generative AI: Assessing the Role of LLMs*. Springer Nature Switzerland, 2024, pp. 129–148.

[33] S. Abualhaija, M. Ceci, N. Sannier, D. Bianculli, L. C. Briand, D. A. Zetzsche, and M. Bodellini, "AI-enabled regulatory change analysis of legal requirements," in *Proceedings of RE 2024*. IEEE, 2024, pp. 5–17.

[34] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, and C. Berger, *Requirements Engineering Using Generative AI: Prompts and Prompting Patterns*. Springer, 2024, pp. 109–127.

[35] S. Lubos, A. Felfernig, T. N. T. Tran, D. Garber, M. E. Mansi, S. P. Erdeniz, and V. Le, "Leveraging llms for the quality assurance of software requirements," in *Proceedings of RE 2024*. IEEE, 2024, pp. 389–397.

[36] T. B. Brown and others, "Language models are few-shot learners," in *Proceedings of NeurIPS 2020*, 2020.

[37] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proceedings of EACL 2021*. ACL, 2021, pp. 255–269.

[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of NeurIPS 2022*, 2022.

[39] The European Parliament and the Council of the European Union, "Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)," 12 2023.

[40] P. de Hert, V. Papakonstantinou, G. Malgieri, L. Beslay, and I. Sánchez, "The right to data portability in the GDPR: towards user-centric interoperability of digital services," *Comput. Law Secur. Rev.*, vol. 34, no. 2, pp. 193–203, 2018.

[41] Article 29 Data Protection Working Party, "Opinion 02/2013 on apps on smart devices," 2013. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp202_en.pdf

[42] European Data Protection Board, "Guidelines 01/2022 on data subject right - right of access," 2022. [Online]. Available: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-012022-data-subject-rights-right-access_en

[43] European Court of Justice, "Joined cases c-141/12 and c-372/12: Judgment of the court (third chamber) of 17 july 2014; ecli:eu:c:2023:369; celex:62012ca0141," 2014.

[44] ——, "Judgment of the court (first chamber) of 4 may 2023; case c-487/21; ecli:eu:c:2023:369; celex:62021cj0487," 2023.

[45] F. Sovrano, M. Lognoul, and A. Bacchelli, "An empirical study on compliance with ranking transparency in the software documentation of EU online platforms," in *Proceedings of ICSE-SEIS'2024*. ACM, 2024, pp. 46–56.

[46] S. Shin and Y. Kim, "Enhancing graph of thought: Enhancing prompts with LLM rationales and dynamic temperature control," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=l32IrJtpOP

[47] S. Sanyal, T. Xiao, J. Liu, W. Wang, and X. Ren, "Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 10 361–10 386.

[48] E. M. Smith, O. Hsu, R. Qian, S. Roller, Y. Boureau, and J. Weston, "Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents," in *Proceedings of ConvAI@ACL 2022*. ACL, 2022, pp. 77–97.

[49] C. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of EMNLP 2016,*. ACL, 2016, pp. 2122–2132.

[50] P. Manakul, A. Liusie, and M. J. F. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of EMNLP 2023*. ACL, 2023, pp. 9004–9017.

[51] L. Zheng and others, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Proceedings of NeurIPS' 2023*, 2023.

[52] X. Hu, G. Li, X. Xia, D. Lo, and Z. Jin, "Deep code comment generation with hybrid lexical and syntactical information," *Empirical Software Engineering*, vol. 25, pp. 2179–2217, 2020.

[53] L. Kuang, C. Zhou, and X. Yang, "Code comment generation based on graph neural network enhanced transformer model for code understanding in open-source software ecosystems," *Autom. Softw. Eng.*, vol. 29, no. 2, p. 43, 2022.

[54] S. Ezzini, S. Abualhaija, C. Arora, and M. Sabetzadeh, "AI-based question answering assistance for analyzing natural-language requirements," in *Proceedings of ICSE 2023*. IEEE, 2023, pp. 1277–1289.

[55] S. Xu, L. Pang, M. Yu, F. Meng, H. Shen, X. Cheng, and J. Zhou, "Unsupervised information refinement training of large language models for retrieval-augmented generation," in *Proceedings of ACL 2024*. ACL, 2024, pp. 133–145.

[56] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL 2002*. ACL, 2002, pp. 311–318.

[57] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. ACL, 07 2004, pp. 74–81.

[58] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of IEEvaluation 2005*. ACL, 2005, pp. 65–72.

[59] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *Proceedings of ICLR 2020*. OpenReview.net, 2020.

[60] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-free? assessing the reliability of leading AI legal research tools," *CoRR*, vol. abs/2405.20362, 2024.

[61] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, "Large legal fictions: Profiling legal hallucinations in large language models," *Journal of Legal Analysis*, vol. 16, no. 1, pp. 64–93, 2024.

[62] L. Humphreys, C. Santos, L. Di Caro, G. Boella, L. Van Der Torre, and L. Robaldo, "Mapping recitals to normative provisions in eu legislation to assist legal interpretation," in *Legal Knowledge and Information Systems*. IOS Press, 2015, pp. 41–49.

[63] L. Chen, M. Zaharia, and J. Zou, "How Is ChatGPT's Behavior Changing Over Time?" *Harvard Data Science Review*, vol. 6, no. 2, 2024.

[64] J. Bhatia, M. C. Evans, and T. D. Breaux, "Identifying incompleteness in privacy policy goals using semantic frames," *Requir. Eng.*, vol. 24, no. 3, pp. 291–313, 2019.

[65] O. Dieste and N. J. Juzgado, "Systematic review and aggregation of empirical studies on elicitation techniques," *IEEE Trans. Software Eng.*, vol. 37, no. 2, pp. 283–304, 2011.

[66] A. G. Sutcliffe and P. Sawyer, "Requirements elicitation: Towards the unknown unknowns," in *Proceedings of RE 2013*. IEEE, 2013, pp. 92–104.

[67] T. D. Breaux and A. I. Antón, "Analyzing regulatory rules for privacy and security requirements," *IEEE Trans. Software Eng.*, vol. 34, no. 1, pp. 5–20, 2008.

[68] A. K. Massey, P. N. Otto, and A. I. Antón, "Aligning requirements with HIPAA in the itrust system," in *Proceedings of RE 2008*. IEEE, 2008, pp. 335–336.

[69] K. Kolthoff, C. Bartelt, S. P. Ponzetto, and K. Schneider, "Self-elicitation of requirements with automated gui prototyping," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 2024, pp. 2354–2357.

[70] J. Wei, A.-L. Courbis, T. Lambolais, B. Xu, P. L. Bernard, G. Dray, and W. Maalej, "Getting inspiration for feature elicitation: App store- vs. llm-based approach," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE '24. Association for Computing Machinery, 2024, pp. 857–869.

[71] S. Ren, H. Nakagawa, and T. Tsuchiya, "Combining prompts with examples to enhance llm-based requirement elicitation," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2024, pp. 1376–1381.

[72] M. R. Tabassum, M. J. Ritchie, S. Mustafiz, and J. Kienzle, "Using llms for use case modelling of iot systems: An experience report," in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, 2024, pp. 611–619.

[73] G. A. Morales, P. K. C, S. Jahan, M. B. Hosseini, and R. Slavin, "A large language model approach to code and privacy policy alignment," in *Proceedings of SANER 2024*. IEEE, 2024, pp. 79–90.

[74] D. Rodriguez, I. Yang, J. M. Del Alamo, and N. Sadeh, "Large language models: a new approach for privacy policy analysis at scale," *Computing*, vol. 106, no. 12, pp. 3879–3903, 2024.

[75] J. Ioannidis, J. Harper, M. S. Quah, and D. Hunter, "Gracenote. ai: Legal generative ai for regulatory compliance," in *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)*, 2023.

[76] K. Ronanki, C. Berger, and J. Horkoff, "Investigating chatgpt's potential to assist in requirements elicitation processes," in *Proceedings of SEAA 2023*. IEEE, 2023, pp. 354–361.