2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA) | 979-8-3503-8674-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/TPS-ISA62245.2024.00021

PrivComp-KG: Leveraging KG and LLM for Compliance Verification

Leon Garza
Computer Science
The University of Texas at El Paso
El Paso, TX, USA
lgarza3@miners.utep.edu

Lavanya Elluri
Computer Information Systems
Texas A&M University Central Texas
Killeen, TX, USA
elluri@tamuct.edu

Aritran Piplai
Computer Science
The University of Texas at El Paso
El Paso, TX, USA
apiplai@utep.edu

Anantaa Kotal
Computer Science
The University of of Texas at El Paso
El Paso, TX, USA
akotal@utep.edu

Deepti Gupta
Computer Information Systems
Texas A&M University Central Texas
Killeen, TX, USA
d.gupta@tamuct.edu

Anupam Joshi
CSEE
University of Maryland Baltimore County
Baltimore, MD, USA
joshi@umbc.edu

***Abstract*—Regulatory documents are complex and lengthy, making full compliance a challenging task for businesses. Similarly, privacy policies provided by vendors frequently fall short of the necessary legal standards due to insufficient detail. To address these issues, we propose a solution that leverages a Large Language Model (LLM) in combination with Semantic Web technology. This approach aims to clarify regulatory requirements and ensure that organizations' privacy policies align with the relevant legal frameworks, ultimately simplifying the compliance process, reducing privacy risks, and improving efficiency. In this paper, we introduce a novel tool, the Privacy Policy Compliance Verification Knowledge Graph, referred to as PrivComp-KG. PrivComp-KG is designed to efficiently store and retrieve comprehensive information related to privacy policies, regulatory frameworks, and domain-specific legal knowledge. By utilizing LLM and Retrieval Augmented Generation (RAG), we can accurately identify relevant sections in privacy policies and map them to the corresponding regulatory rules. Our LLM-based retrieval system has demonstrated a high level of accuracy, achieving a correctness score of 0.9, outperforming other models in privacy policy analysis. The extracted information from individual privacy policies is then integrated into the PrivComp-KG. By combining this data with contextual domain knowledge and regulatory rules, PrivComp-KG can be queried to assess each vendor's compliance with applicable regulations. We demonstrate the practical utility of PrivComp-KG by verifying the compliance of privacy policies across various organizations. This approach not only helps policy writers better understand legal requirements but also enables them to identify gaps in existing policies and update them in response to evolving regulations.**

***Index Terms*—Privacy Policy, Policy Compliance, Large Language Model, Knowledge Graph.**

I. Introduction

As data collection from diverse sources around the world continues to grow exponentially, concerns about how this data is managed and protected are becoming increasingly significant. Large amounts of data is being gathered almost constantly, ranging from individual interactions on social media to sensor readings from phones. This data is highly valuable, offering deep insights into consumer behaviors, market trends, and societal patterns. Businesses leverage this information to tailor their products and services, streamline operations, and gain a competitive edge in the marketplace. At the same time, researchers analyze vast datasets to drive scientific breakthroughs, improve healthcare outcomes, and tackle societal challenges. This increasing reliance on data, however, amplifies the need for robust data privacy measures. Ensuring that privacy is maintained while still enabling the valuable use of data is a critical challenge.

There is potential for misuse and unauthorized dissemination of consumers' private information by organizations collecting their data. To address this, numerous data protection regulations, such as the European Union's General Data Protection Regulation (GDPR) [1], Payment Card Industry Data Security Standard (PCI DSS) [2], and the California Consumer Privacy Protection Act (CCPA) [3], [4], have been established in response to public apprehension. Privacy Policy regulations impose strict rules on collecting, using, and managing personal data. These regulations impose strict guidelines on how companies collect, use, and manage personal data. Key principles include minimizing data collection, limiting usage to specific purposes, obtaining user consent, and ensuring data accuracy, security, and accountability. The GDPR, for example, sets a strict set of privacy and data protection rules for companies accessing users' data under the jurisdiction of the European Union (EU). Known for setting some of the strictest data privacy and security standards worldwide, the GDPR is a model for properly using personal data, focusing on safeguarding and legally processing individuals' information. As a result, businesses must reassess how they handle data, ensuring their approaches comply with GDPR guidelines to protect people's privacy while maximizing benefits with data in this digital era. Furthermore, not adhering to this regulations not only makes the organization more susceptible to data breaches, but also holds them liable to pay huge penalties.

In May 2023, the Irish Data Protection Commission (DPC)

979-8-3503-8674-5/24/$31.00 ©2024 IEEE
DOI 10.1109/TPS-ISA62245.2024.00021

made a major move in the history of the GDPR by fining the American tech company Meta a record €1.2 billion [5]. This fine was the highest ever because Meta moved the personal data of European users to the United States without ensuring enough protection for this data. This step by the DPC is a crucial moment in data protection law, highlighting how seriously rules about international data transfer are enforced under the GDPR. In July 2021, Amazon Europe Core SARL was fined €746 million by the Luxembourg National Commission for Data Protection (CNDP), marking the most significant penalty for violating the GDPR. This action followed a complaint from 10,000 individuals, organized by the French privacy group La Quadrature du Net, concerning Amazon's data processing practices. The CNDP's investigation revealed that Amazon's advertising targeting system operated without obtaining proper consent, contravening GDPR's stringent consent requirements, which demand clear communication and detailed explanations of personal data use, purpose, and usage. [6] In September 2023, TikTok faced a significant penalty from the Irish DPC, receiving a fine of EUR 345 million. This event marked one of the most considerable GDPR fines imposed on a social media platform, particularly highlighting issues around protecting children's data privacy. The decision underscored the critical need for tech companies to prioritize the safety of young users online. [7] These instances underscores the critical need for businesses to understand privacy policy regulations and ensure that their data privacy policies are consistent with the regulations.

Businesses must thoroughly understand the nature, scope, and purpose of the data they collect, process, and store. Furthermore, this information must be precisely recorded in a privacy policy document that's easy for users to access and understand. Writing a comprehensive privacy policy document is critical to building trust between data collectors and consumers. Developing a privacy policy that meets the extensive requirements of policy regulations is a significant challenge for companies, primarily because of the complexity of the regulation's rules. Privacy policies are short and concise for user ease while complying with all relevant sections of the regulatory document.

Furthermore, the legal landscape of data protection regulations is dynamic and evolving. Regulations, such as GDPR, set a high privacy and data protection standard, yet it's subject to interpretation and ongoing adjustments by regulatory authorities. This ever-changing landscape necessitates that companies stay flexible and constantly monitor legal updates to ensure their privacy policies and practices align with compliance requirements. Additionally, the global reach of data sharing complicates compliance efforts. Companies must navigate and identify all relevant privacy laws and comply. Given these challenges, the motivation behind this work is to provide a solution that helps privacy policy writers efficiently develop policies that are not only comprehensive but also in full compliance with the latest regulatory documents. By streamlining the process of policy creation and ensuring alignment with evolving legal standards, this approach aims to reduce the burden on companies and enhance their ability to protect data in a rapidly changing digital landscape.

In this paper, we propose using a Large Language Model (LLM) and Semantic Web technology to provide clear interpretations of regulatory requirements and ensure that an organization's privacy policies align with said regulations. The novel framework enables policy writers to identify relevant regulatory rules, detect shortcomings in existing policies, and effectively address compliance requirements. In this work, we develop a Privacy Policy Compliance Verification Knowledge Graph, PrivComp-KG, that is designed to collect and maintain information regarding policy and regulatory documents, and encapsulate domain knowledge. The inference rule engine is used to reason over the privacy policies and regulatory documents to verify compliance. The query engine is utilised to effectively gain this insight from the PrivComp-KG and can be utilised by policy writers to identify any gaps in their existing policies. To efficiently populate the PrivComp-KG with privacy policy documents, as well as their relevance to regulatory sections, we use Retrieval Augmented Generation (RAG) to assess privacy policies' alignment with GDPR articles. This avoids the need for constant model fine-tuning with evolving privacy laws. RAG helps generate responses by utilizing chunks of GDPR articles, allowing us to identify specific segments that match privacy policies, ensuring a dynamic and comprehensive approach without requiring continual model updates for each policy change. We demonstrate the utility of the PrivComp-KG, by verifying compliance of privacy policy documents for various organizations in the OPP-115 dataset [8].

The structure of the paper is organized as follows. In Section II we describe relevant work in contemporary GDPR compliance efforts and LLM RAG methods for text retrieval. In Section III, we describe our framework leveraging LLMs for RAG to align GDPR sections with privacy policies. This section also discusses the development of PrivComp-KG and illustrates how data retrieved from vendor policies are incorporated into the knowledge graph. In Section IV we detail the utilisation of our framework to verify compliance of privacy policies in the OPP-115 dataset. Furthermore we describe the results from evaluation of the knowledge extraction and KG creation methods. The final section summarizes these discoveries and outlines prospective research avenues.

II. Background and Related Work

As data collection increases, public concern over privacy risks has intensified, leading to the implementation of stringent privacy regulations like the GDPR. These regulations are crucial for safeguarding privacy, yet ensuring automatic compliance remains a complex challenge. Privacy policy compliance has been extensively studied, with several key contributions highlighting the challenges and frameworks necessary for effective adherence to regulations. For instance, Barth et al. introduced the concept of contextual integrity, emphasizing the importance of aligning privacy policies with the expectations and norms of different contexts to ensure compliance [9].

Antón et al. proposed a privacy goal taxonomy that can be employed to analyze and ensure compliance with privacy regulations in web-based applications [10].

The General Data Protection Regulation (GDPR) has been particularly influential, with De Hert et al. discussing the challenges organizations face in complying with its stringent requirements [11]. Zimmeck et al. introduced the Mobile App Privacy System (MAPS) [12], a tool designed for large-scale privacy compliance analysis of Android apps. The system compares the actual privacy practices of apps, determined through code analysis, with their stated privacy policies, uncovering widespread potential non-compliance issues across over a million apps in the Google Play Store. Korba et al. [13] present a method for discovering private data in collaborative environments by employing named entity recognition (NER) and relation extraction (RE) techniques, using supervised machine learning to identify and manage personally identifiable information (PII) within semi-structured and unstructured documents, ultimately aiming to support privacy compliance in organizations. Srinath et al. [14] introduce the PrivaSeer Corpus, a large-scale collection of over a million website privacy policies, and demonstrate its use in pretraining PrivBERT, a privacy-focused language model that achieves state-of-the-art results in data practice classification and question answering tasks. The PrivBERT language model has been used to identify and visualize data practices within privacy policies by matching policy excerpts with predefined descriptions.

In recent research [15], the concept of "Data Capsule" is introduced, automating compliance checks against privacy regulations in data processing. Individual data is associated with specific policies, ensuring adherence through residual policies and a new algorithm for effective policy derivation. This system advances individual privacy protection, albeit focusing solely on data subject rights. Another study [16] proposes an approach to assess privacy policy alignment with GDPR Article 13 standards. By manually selecting 304 policies and developing a labeling system, the authors identify compliance issues, creating a web tool named AutoCompliance to simplify policy comprehension. However, this study overlooks broader GDPR coverage. In privacy policy research [17], the impact of GDPR on over 6,000 policies is analyzed, indicating significant revisions post-GDPR, particularly in EU policies. User experience improvements are noted, but confidence in vendor compliance remains uncertain. Leveraging ontology and text extraction techniques [18], vendors automate privacy policy compliance efforts, streamlining data protection measures. These advancements signify progress in managing privacy constraints, but comprehensive compliance assurance remains a challenge.

In our earlier research studies, we developed a foundational compliance knowledge graph to include various regulations and incorporated a selection of vendor privacy policy descriptions into the ontology without directly linking them to specific GDPR chapters or sections. Also, we correlated these documents with Cloud Security Alliance (CSA) controls to bridge gaps. In our past research [19], [20], we identified relevant sections by extracting keywords and entities from the glossaries or appendices of regulations. We then identified the semantically similar keywords associated with GDPR regulation from the vendor privacy policies. Further, we checked for the semantic similarity between the summaries of the entire GDPR and the privacy policy document using a generic BERT abstractive summarize [21], [22]. In another research work, we have incorporated the National Institute of Standards and Technology (NIST) 8228 [23] risk mitigation areas into the knowledge graph. [24], [25].

Expanding upon this groundwork, our current research seeks to align the extracted GDPR articles from vendor privacy policies, pinpointing any previously overlooked articles. This effort is designed to assist vendors in refining their policy documents. Given the extensive nature of these regulations, our focus is primarily on GDPR compliance, recognizing its significance for vendors handling data from EU users. The traditional approach often necessitates manual review to ensure compliance. However, our methodology proposes a more efficient solution for identifying and addressing gaps in vendor privacy policies, reducing the need for human intervention.

Although mostly statistical and rule-based methods have been used for information extraction [26], [27], recent advancements in LLMs have opened up new methodologies.LLMs have been extensively used for information extraction across various domains. For information extraction tasks, LLMs have been utilized for generating structured entities and relationships circumventing the need to use supervised models [28]–[30]. However, to find more success in specific domains, LLMs have been fine-tuned to perform the task of information extraction. For example, in the case of scientific data extraction [31] and agriculture data extraction [32], fine-tuned LLMs have been used. LLMs have also been used for improving annotations in the medical domain, by periodically fine-tuning based on human feedback [33]. In the domain of cybersecurity, LLMs have also been used for information combination and extraction [34], [35]. However, training and fine-tuning an LLM is computationally expensive and there is little guarantee that the model will not suffer from hallucinations. In our approach, we have utilized the power of RAG to limit the possibilities of hallucinations and avoid the cost of fine-tuning for our specific application scenario.

III. Methodology

The Privacy Policy Compliance Verification Knowledge Graph (PrivComp-KG) formalizes GDPR rules and guidelines using Semantic Web technologies. It facilitates automated compliance checking, enhances transparency, and supports granular consent management. The Knowledge Graph allows for swift cross-referencing of regulatory requirements and vendor privacy policies, enabling efficient management of vendor data and adherence to regulations. Compliance inference using SWRL (Semantic Web Rule Language) rules [36] enriches the understanding of privacy policies, facilitating dynamic compliance and reasoning over gaps in policies. The PrivComp-KG is populated with relevant privacy policy

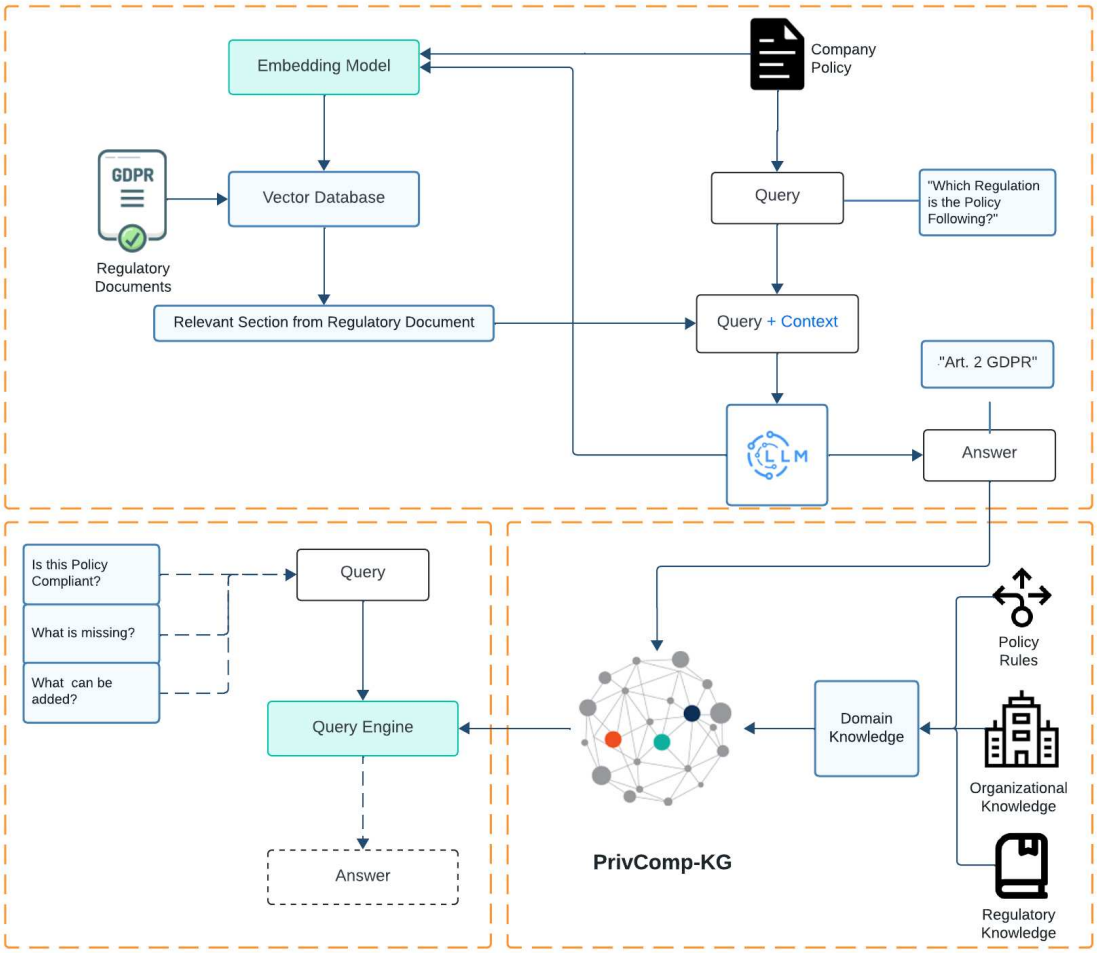

Fig. 1: Overall Framework for Building and Querying GDPR Vendor Policy Management Knowledge Graph

properties by leveraging LLMs to assess privacy policies' alignment with regulatory articles, such as GDPR. We employ Retrieval-Augmented Generation (RAG) to mitigate LLMs' tendency for hallucinations when confronted with unfamiliar queries, dynamically generating responses without continual fine-tuning. The overall framework for PrivComp-KG creation and population using LLMs and end user querying is demonstrated in Figure 1.

A. *Knowledge Extraction using LLM*

LLMs are increasingly valued for their deep understanding of natural language. The vector representations of each piece of text written in natural language, have a deeper contextual meaning in the scope of LLMs.

To understand privacy policies, we harness the capabilities of LLMs to assess the alignment of a privacy policy with GDPR articles. Previous efforts have focused on fine-tuning LLMs to identify similar GDPR entities corresponding to a specific privacy policy [21]. However, both LLMs in general and fine-tuned LLMs specifically are prone to hallucinations when confronted with queries relating to unfamiliar domains. An increasingly popular method to address this issue is RAG. We opt for RAG due to the dynamic nature of the domain. Given that data privacy laws can be evolving, characterized by the emergence of new regulations, RAG aids in response generation without the need for continual model fine-tuning for every update in privacy laws and regulations.

The core components of an LLM consist of (i) a query or prompt, denoted as P, and (ii) a response, represented by R. In the context of RAG-enabled LLMs, the generation of R relies on a set of documents, denoted as D. We utilize RAG not only to produce responses for prompts inquiring about the relationship between a privacy policy and a GDPR article but also to identify the specific segments of a GDPR article that align with a privacy policy.

We segment each GDPR article into chunks and integrate them into a vector store. Each segment of the GDPR article is assigned a representation within the vector store. Our

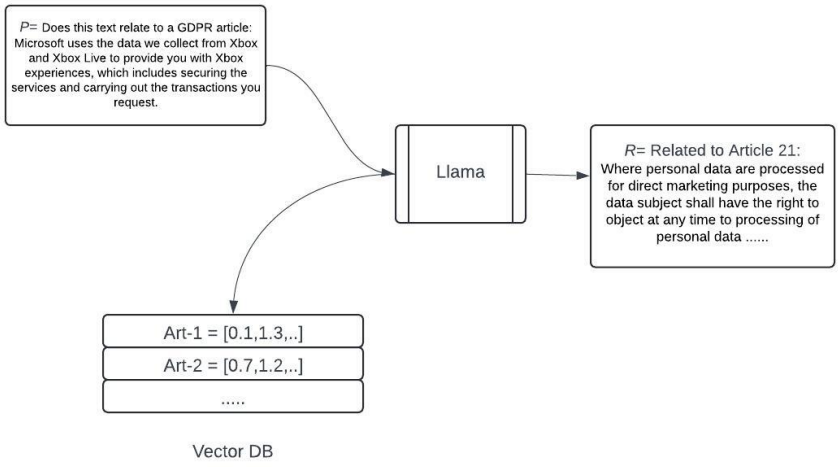

Fig. 2: Example of our model's response generation on a privacy policy sample

vector store contains 430 chunks of GDPR articles, with each article further divided into multiple segments. Typically, each paragraph of a GDPR article corresponds to one vector within our vector store. An example of our prompt P is illustrated in Figure 2.

As part of the metadata for the chunks inserted into the vector store, we include the specific GDPR article from which each chunk was extracted. A typical P for this system might be "Which GDPR article does this privacy policy relate to?" followed by the privacy policy itself. The retriever then (i) generates R based on the GDPR articles and (ii) provides a list of articles used to synthesize R, along with their corresponding similarity scores.

In Figure 2, we observe an example of the model's performance using an excerpt from Microsoft's XBOX privacy policy. Our vector store is constructed by inserting chunks of GDPR articles, which are then utilized by our LLM (LLama-7B) to generate R. In the example, we highlight Article 21, which achieved the highest similarity score with P. During our experiments, we establish a threshold for the similarity score and list all articles used in generating R.

This process creates a comprehensive list of articles that correlate with a given privacy policy. Our system supports dynamic updates to privacy policies, as model retraining is unnecessary when policies are amended or added. The knowledge derived from privacy policies and their corresponding articles can be incorporated into our PrivComp-KG. In the following sections, we elaborate on how the reasoning capabilities of a knowledge graph can be leveraged to derive valuable insights from privacy policies and the associated GDPR articles.

B. PrivComp-KG: Privacy Policy Compliance Verification Knowledge Graph

Leveraging Semantic Web technologies can streamline regulatory compliance. Using the Semantic Web for privacy policy compliance offers advantages such as standardization and machine readability. It enables automated compliance checking by representing policies in a format that software tools can understand and analyze. This approach enhances transparency by allowing policy writers to easily comprehend data privacy requirements and gaps in their existing policies. Furthermore, the semantic representations facilitate granular consent management, tailoring policies to specific regulations.

We utilized the semantic web languages Resource Description Framework (RDF) [37] and Web Ontology Language (OWL) [38] to capture and formalize the rules and guidelines outlined in GDPR and Vendor policy documents. We developed the novel Privacy Compliance Verification KG, PrivComp-KG. It is designed to be in the public domain and can be adopted quickly and easily by vendors who are seeking to adhere to these regulations. The ontology is also platform-independent and can be integrated with the latest data protection regulations and many other data regulation entities.

RDF enhances the structuring of knowledge on the web, simplifying the retrieval of domain-specific information for vendors. PrivComp-KG is integrated with our existing Reference Document Knowledge Graph [21], allowing for the swift and effective cross-referencing of regulatory requirements and vendor privacy policies. As illustrated in Figure 3, this high-level knowledge graph manages GDPR rules and vendor policy extracted results. This knowledge graph is specifically designed to accommodate any applicable regulations for various vendors or companies based on the types of data they collect. Our focus on GDPR, a pivotal regulation with extensive stipulations, forms a critical component of

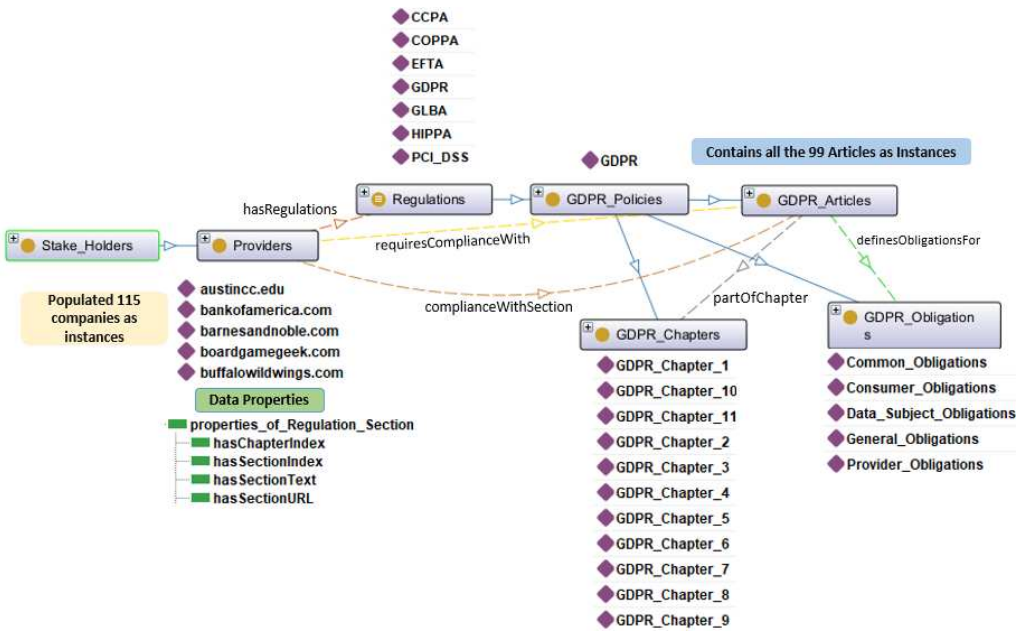

Fig. 3: PrivComp-KG: Privacy Policy Compliance Verification Knowledge Graph along with Instances of the classes

this research methodology. Utilizing Protege software [39], [40], a free, open-source platform for ontology editing and reasoning, we constructed and managed our ontology. This structured and standardized approach not only facilitates the management of vendor data but also ensures adherence to regulations, thereby safeguarding customer data and privacy. Our knowledge graph is hosted in a public space, providing accessible user interaction.

In our earlier research, described in [21], we could only retrieve high-level entities and didn't examine the specific rules or articles related to privacy policies. Additionally, our previous version of the knowledge graph categorized most GDPR rules as classes, which made it challenging to compare results across different sections of the GDPR. Structuring these as instances improved our ability to identify what was missing quickly or had been extracted. We used the insights from the Section III-A to update the data properties of the *Provider* class. With this research, we've successfully refined our knowledge graph to manage and search GDPR rules flexibly and compare them with those extracted from vendor policy documents. We designed this knowledge graph to include a Regulations class that branches into various regulations like PCI-DSS, HIPAA, CCPA, etc. This study extracted relevant chapters, articles, and obligations from the GDPR and stored them as instances in *GDPR_Chapter*, *GDPR_Articles*, and *GDPR_Obligations*.

C. Regulatory Obligations:

Beyond the curation of Provider Privacy Policies and Policy Regulations, the key insight in PrivComp-KG is drawn from the Regulatory Obligations instances. Every regulation describes based on the role of the data actor, the specific set of rules that apply to the data actor. For example, a provider, i.e. an organization offering a specific service, may collect data to support that service. However, GDPR clearly states in its provider guidelines, the specific measures that the providers need to take to protect data privacy. Additionally, some general rules with regards to the ethics and responsibility of data collection apply to the providers. This knowledge about the specific role and application of GDPR sections is encoded in the *GDPR_Obligations* class. Specifically, there are 5 instances of the *GDPR_Obligations*, each describing a specific role:

- *Consumer_Obligations* Consumers must inform the supervisory authority and the data subject about any personal data breaches. Additionally, the consumer must conduct a Data Protection Impact Assessment (DPIA), consult with the supervisory authority before processing if the DPIA indicates a high risk, and appoint a Data Protection Officer when processing personal data on a large scale.
- *Common_Obligations* Rules that apply to both consumers and providers, who are responsible for ensuring compliance.
- *Data_Subject_Obligations* Consumers must inform data subjects about the duration for which, or why, data will be retained upon collection. Consequently, if data subjects request the removal of their data, and it is no longer necessary for the purposes for which it was collected, it must be erased.
- *General_Obligations* GDPR mentions generic rules that are not specific to any role but apply to all the data processing activities in general.

- *Provider_Obligations* Provider is primarily responsible for assisting consumers in the event of data breaches and processing data in accordance with consumer directives. Additionally, the Provider must maintain comprehensive records of all data processing activities and ensure robust data security measures are in place to protect consumer information.

D. Compliance and Regulatory Properties:

The PrivComp-KG supports object properties that links classes to support compliance reasoning and verification.

- hasRegulation: Regulations identified in vendor privacy policies are stored as instances within the *Regulation* class and linked to the *Providers* class.
- compliesWithSection: Using the knowledge extraction tool, as described in Section III-A, the regulatory articles that individual provider privacy policies complies with are identified. This knowledge is populated into the PrivComp-KG using the "compliesWithSection" relation between *Providers* and *GDPR_Articles*.
- requriesComplianceWith: Every vendor handling EU user data must adhere to relevant GDPR requirements. To enforce this, we link the *Providers* class to the *GDPR_Chapters* through this object property.
- partOfChapter: Since a chapter can encompass multiple articles, we connect the *GDPR_Articles* to the *GDPR_Chapters* class to reflect this relationship.

Additionally, to organize the findings from privacy policy analyses, we established several data properties:

- hasChapterIndex: This property stores the indices of chapters identified in a policy document.
- hasSectionIndex, hasSectionText, and hasSectionURL: These properties are essential for recording the sections extracted from the documents, including descriptions and URLs for easy reference.

E. Compliance Inference

SWRL rules enhance reasoning capabilities in Semantic Web applications by allowing the specification of logical rules that define relationships and infer new information from existing data. The PrivComp-KG supports SWRL rules that infers necessary rules from regulatory articles based on the role of the data actor and the role based obligations, as described in Section III-C.

For example, for the privacy policy of a provider that wishes to collect and utilise data, the following SWRL rules can infer the obligatory rules from GDPR.

Listing 1: SWRL Rule 1

```
S1: Cloud_Providers(?cloud_provider) ^
    GDPR_Articles(?gdpr_article)
^ definesObligationsFor(?gdpr_article,
    Provider_Obligations)
-> requiresComplianceWith(?cloud_provider,
    ?gdpr_article)
```

Listing 2: SWRL Rule 2

```
S2: Cloud_Providers(?cloud_provider) ^
    GDPR_Articles(?gdpr_article)
^ definesObligationsFor(?gdpr_article,
    Common_Obligations)
-> requiresComplianceWith(?cloud_provider,
    ?gdpr_article)
```

By applying these SWRL rules, PrivComp-KG can derive detect inconsistencies in existing policies, and make logical deductions about compliance. This reasoning process enriches the understanding and interpretation of privacy policy data, facilitating a dynamic approach towards privacy policy compliance, more advanced semantic querying and data integration. After processing with the reasoner, users can efficiently compare the required rules against those extracted, updating any lacking areas in the privacy policy to ensure it is current and compliant.

IV. Experimental Results

A. Dataset

The OPP-115 dataset [8] is a comprehensive collection of privacy policies from various online platforms, consisting of over 115 privacy policies. It encompasses a wide range of websites and services, including social media platforms, e-commerce sites, and mobile applications. The dataset is structured and annotated, making it suitable for privacy policy analysis and evaluating language models. It includes the categorisation of data collection, usage, sharing, and retention practices outlined in the privacy policies. In this work, we use the OPP-115 dataset to demonstrate the utility of our model and evaluate the performance of the LLMs. Our LLM based method is used to identify relevant regulatory articles for each provider in the dataset and then populated into PrivComp-KG. The inferential engine in PrivComp-KG reasons over the gaps in these provider policies. The results are made available to end users using a query engine. The results from our evaluation methods and query engine are described in subsequent sections.

B. Evaluation of LLM-guided extraction

Threshold Used	Correctness Score
0.9	0.66
1.0	0.74
1.1	0.82
1.2	0.84
1.3	0.89
1.4	0.88
1.5	0.9

TABLE I: Correctness score for each threshold

In our experiments, we employ the Llama-7B large language model (LLM) alongside *chromaDB* as our vector store to evaluate the performance of a Retrieval-Augmented Generation (RAG) enabled LLM in analyzing privacy policies. The knowledge extracted by the RAG-LLM is specifically targeted at GDPR articles corresponding to various sections of privacy policies. We assess this knowledge through two key metrics:

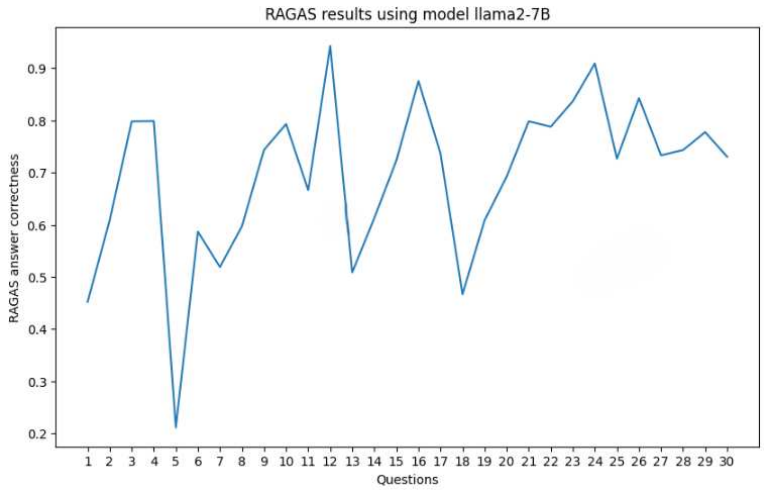

Fig. 4: RAGAS [41] score evaluating the quality of privacy policy "Answer" generation through our LLM, in comparison with human-generated "Answers". A score of '1' means perfect alignment.

(a) GDPR articles with which the privacy policy currently complies (b) GDPR articles with which the privacy policy needs to comply

Fig. 5: Compliance results for Lids.com from PrivComp-KG

Classification	Correctness Score
SVM+TFIDF	0.8
SVM+Word2Vec	0.75
PrivBERT	0.89
LLAMA + Threhold=1.5	0.9

TABLE II: Comparison of Correctness Score for Privacy Policy Classification Models

(i) the quality of the responses generated by the RAG-LLM and (ii) the accuracy of the GDPR article numbers retrieved by the LLM.

Given that RAG plays a crucial role in the LLM's functionality, we first evaluate the effectiveness of the responses (R) generated by our model in relation to the input privacy policy (P). One of the common metrics used for evaluating RAG-generated responses is RAGAS [41]. In our evaluation, we compare the responses generated by our model with those generated by human experts. As illustrated in Figure 4, the correctness scores for all questions are presented, highlighting that questions requiring highly specific information about the privacy policies tend to receive lower scores (<0.5). This is expected, as our LLM primarily accesses GDPR articles, allowing it to perform well in questions regarding the alignment of privacy policies with GDPR articles, but it struggles with more detailed, policy-specific inquiries.

For the second metric, we utilize the OPP-15 dataset, which maps privacy policies to GDPR articles through an intermediate categorization. We leverage this dataset to assess the accuracy of the articles retrieved during R generation. To ensure the reliability of our evaluation, we preprocess the dataset to remove sections of privacy policies that are too small to be meaningful. We then apply a similarity score function provided by RAG, setting a threshold to select only those chunks that meet our criteria. Table I presents the accuracy of correctness for various thresholds. Furthermore, in Table II, we compare our model's performance with other Privacy Policy Classification models, such as a Support Vector

Machine classifier with TF-IDF and Word2Vec embeddings, as well as the PrivBERT model. The results indicate that the Llama-7B model, when used with a threshold of 1.5, significantly outperforms the other models, achieving an impressive correctness score of 0.9.

This superior performance underscores the effectiveness of our model in accurately aligning privacy policies with relevant GDPR articles, making it a powerful tool for privacy policy analysis and compliance verification.

C. Knowledge Graph Inferences and Queries

The PrivComp-KG is designed to streamline the process of privacy policy analysis by integrating the provider privacy policies from the OPP-115 dataset and mapping them to relevant GDPR articles. Through the use of the "compliesWithSection" object property, the identified regulatory sections of relevance to the privacy policies are automatically linked within PrivComp-KG. Additionally, the system utilizes a reasoner to dynamically update the GDPR articles that are mandatory for each provider, employing the "requiresComplianceWith" object property to ensure comprehensive coverage.

As demonstrated in Fig 5, the PrivComp-KG is particularly effective in identifying compliance and gaps in privacy policies. For the provider "Lids.com," the LLM results are used to populate the PrivComp-KG with the GDPR articles that "Lids.com" currently complies with, as shown in Fig 5 (a). Moreover, leveraging the knowledge graph (KG) reasoner, PrivComp-KG also identifies the GDPR articles that "Lids.com" is required to comply with, as depicted in Fig 5 (b). This dual approach of assessing both compliance and required compliance provides a comprehensive view of the provider's obligations under GDPR.

This functionality is particularly valuable for policy writers at organizations like "Lids.com," as it allows them to quickly identify gaps in their existing privacy policies. To query and identify the specific GDPR articles that are missing from a provider's compliance framework, PrivComp-KG offers an efficient SPARQL query:

```
SELECT * WHERE {
{cc:lids.com cc:requiresComplianceWith ?x}
MINUS {cc:lids.com cc:compliesWithSection ?x}
}
```

For "Lids.com," the PrivComp-KG returns 40 GDPR articles that have been identified as mandatory but are currently missing from the existing privacy policy. This critical information enables policy writers to efficiently pinpoint and address the gaps in their privacy policies, ensuring that they meet all necessary regulatory requirements. The effectiveness of PrivComp-KG in automating this process not only saves time but also enhances the accuracy and thoroughness of privacy policy compliance efforts.

V. Conclusion and Future Work

In the digital age, safeguarding data protection and privacy has become paramount. As companies increasingly rely on third-party vendors and service providers for critical functions like data handling, storage, and processing, the need for robust data protection measures has intensified. Entrusting sensitive data to these external partners necessitates rigorous protocols and contractual agreements to ensure the integrity and confidentiality of the information shared. While existing research often focuses on specific sections of regulations, this approach falls short of providing comprehensive support for organizations aiming to maintain full compliance.

In this research, we addressed this gap by leveraging Large Language Models (LLM) and Semantic Web technologies to create a more holistic solution for verifying the compliance of privacy policy documents with policy regulations such as GDPR. Our novel contribution, the Privacy Policy Compliance Verification Knowledge Graph, PrivComp-KG, serves as a dynamic repository for storing and retrieving comprehensive information related to privacy policies. By integrating LLM results with a structured knowledge graph and reasoner, PrivComp-KG offers an effective means of identifying compliance gaps, automatically updating mandatory GDPR articles, and enhancing the overall readability and transparency of privacy policies.

The benefits of this research are multifaceted: it promotes transparency, empowers consumers, strengthens regulatory compliance, and ultimately fosters trust in the digital ecosystem. By making privacy policies more accessible and easier to understand, PrivComp-KG also helps organizations quickly identify and address any deficiencies in their compliance frameworks. Looking ahead, we plan to further enhance PrivComp-KG by incorporating multiple data protection regulations, enabling rapid cross-referencing across a comprehensive legislative framework. This will provide even greater support to companies in their efforts to navigate the complex and evolving landscape of data protection and privacy regulations, ensuring that they remain compliant and trustworthy in an increasingly data-driven world.

Acknowledgment

This work was supported by NSF award #2348147. We express our gratitude to colleagues whose insights and expertise significantly contributed to the research.

References

[1] E. Union. (2024, 04) General data protection regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.
[2] P. S. Council. (2024, 04) Payment card industry mobile payment guidelines. https://www.pcisecuritystandards.org/document_library/.
[3] "California consumer privacy act of 2018," State of California, Tech. Rep., 2018.
[4] "California consumer privacy act of 2020," State of California, Tech. Rep., 2020.
[5] E. D. P. Board. Meta. https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision_en.
[6] N. L. Review. Luxembourg amazon gdpr violations. https://natlawreview.com/article/luxembourg-dpa-fines-amazon-746-million-euros-gdpr-violations.
[7] T. I. Times. Tiktok fined €345m by ireland's data regulator for violating children's privacy. https://www.irishtimes.com/technology/big-tech/2023/09/15/tiktok-fined-345m-by-irelands-data-regulator-for-violating-childrens-privacy/.

[8] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1330–1340.

[9] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum, "Privacy and contextual integrity: Framework and applications," in *2006 IEEE symposium on security and privacy (S&P'06)*. IEEE, 2006, pp. 15–pp.

[10] A. I. Antón, J. B. Earp, and A. Reese, "Analyzing website privacy requirements using a privacy goal taxonomy," in *Proceedings IEEE Joint International Conference on Requirements Engineering*. IEEE, 2002, pp. 23–31.

[11] P. De Hert and V. Papakonstantinou, "The new general data protection regulation: Still a sound system for the protection of individuals?" *Computer law & security review*, vol. 32, no. 2, pp. 179–194, 2016.

[12] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh, "Maps: Scaling privacy compliance analysis to a million apps," *Proceedings on Privacy Enhancing Technologies*, 2019.

[13] L. Korba, Y. Wang, L. Geng, R. Song, G. Yee, A. S. Patrick, S. Buffett, H. Liu, and Y. You, "Private data discovery for privacy compliance in collaborative environments," in *Cooperative Design, Visualization, and Engineering: 5th International Conference, CDVE 2008 Calvià, Mallorca, Spain, September 21-25, 2008 Proceedings 5*. Springer, 2008, pp. 142–150.

[14] M. Srinath, S. Wilson, and C. L. Giles, "Privacy at scale: Introducing the privaseer corpus of web privacy policies," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, 2021.

[15] *Data capsule: A new paradigm for automatic compliance with data privacy regulations*. Springer, 2019.

[16] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang, "Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13," in *Proceedings of the Web Conference 2021*, 2021, pp. 2154–2164.

[17] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the gdpr," *arXiv preprint arXiv:1809.08396*, 2018.

[18] Y.-J. Hu, H.-Y. Guo, and G.-D. Lin, "Semantic enforcement of privacy protection policies via the combination of ontologies and rules," in *2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC 2008)*. IEEE, 2008, pp. 400–407.

[19] L. Elluri, K. P. Joshi, and A. Kotal, "Measuring semantic similarity across eu gdpr regulation and cloud privacy policies," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3963–3978.

[20] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate gdpr and pci dss compliance," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1266–1271.

[21] L. Elluri, S. S. L. Chukkapalli, K. P. Joshi, T. Finin, and A. Joshi, "A bert based approach to measure web services policies compliance with gdpr," *IEEE Access*, vol. 9, pp. 148 004–148 016, 2021.

[22] A. Kotal, L. Elluri, D. Gupta, V. Mandalapu, and A. Joshi, "Privacy-preserving data sharing in agriculture: Enforcing policy rules for secure and confidential data synthesis," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 5519–5528.

[23] K. Boeckl, K. Boeckl, M. Fagan, W. Fisher, N. Lefkovitz, K. N. Megas, E. Nadeau, D. G. O'Rourke, B. Piccarreta, and K. Scarfone, *Considerations for managing Internet of Things (IoT) cybersecurity and privacy risks*. US Department of Commerce, National Institute of Standards and Technology ..., 2019.

[24] K. U. Echenim, L. Elluri, and K. P. Joshi, "Ensuring privacy policy compliance of wearables with iot regulations," in *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2023, pp. 247–256.

[25] K. Boeckl, K. Boeckl, M. Fagan, W. Fisher, N. Lefkovitz, K. N. Megas, E. Nadeau, D. G. O'Rourke, B. Piccarreta, and K. Scarfone, *Considerations for managing Internet of Things (IoT) cybersecurity and privacy risks*. US Department of Commerce, National Institute of Standards and Technology ..., 2019.

[26] A. Pingle, A. Piplai, S. Mittal, A. Joshi, J. Holt, and R. Zak, "Relext: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 879–886.

[27] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, and R. Zak, "Creating cybersecurity knowledge graphs from malware after action reports," *IEEE Access*, vol. 8, pp. 211 691–211 703, 2020.

[28] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, and E. Chen, "Large language models for generative information extraction: A survey," *arXiv preprint arXiv:2312.17617*, 2023.

[29] L. Peng, Z. Wang, F. Yao, Z. Wang, and J. Shang, "Metaie: Distilling a meta model from llm for all kinds of information extraction tasks," *arXiv preprint arXiv:2404.00457*, 2024.

[30] Z. Li, Y. Zeng, Y. Zuo, W. Ren, W. Liu, M. Su, Y. Guo, Y. Liu, X. Li, Z. Hu *et al.*, "Knowcoder: Coding structured knowledge into llms for universal information extraction," *arXiv preprint arXiv:2403.07969*, 2024.

[31] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, 2024.

[32] R. Peng, K. Liu, P. Yang, Z. Yuan, and S. Li, "Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data," *arXiv preprint arXiv:2308.03107*, 2023.

[33] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha *et al.*, "Llms accelerate annotation for medical information extraction," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 82–100.

[34] Z. Liu, J. Shi, and J. F. Buford, "Cyberbench: A multi-task benchmark for evaluating large language models in cybersecurity."

[35] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Piplai, M. Gaur, and S. Rahimi, "Localintel: Generating organizational threat intelligence from global and local cyber knowledge," *arXiv preprint arXiv:2401.10036*, 2024.

[36] A. SWRL, "Semantic web rule language combining owl and ruleml," *W3C Member Submission (May 21, 2004), http://www. w3. org/Submission/SWRL/(last visited March 2011)*, 2004.

[37] W3C. (15 March, 2014) Resource Description Framework. Accessed: August 23, 2023. [Online]. Available: https://www.w3.org/RDF/

[38] W3C. (10 February, 2004) Web Ontology Language. Accessed: August 23, 2023. [Online]. Available: https://www.w3.org/TR/owl-features/

[39] W3C. (2022, March 16) Protégé (software). https://en.wikipedia.org/wiki/Prot%C3%A9g%C3%A9_(software). Accessed: March 20, 2024.

[40] Protege. (2020) Protege tool. http://protege.stanford.edu. Accessed: March 20, 2024.

[41] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *arXiv preprint arXiv:2309.15217*, 2023.