UNIVERSITY OF CALIFORNIA
RIVERSIDE

State of Title IX: A Knowledge Base for Title IX Documentation

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Computer Science

by

Priyanshu Sharma

June 2024

Thesis Committee:
Dr. Yue Dong, Chairperson
Dr. Jiachen Li
Dr. Greg Ver Steeg
Dr. Evangelos Papalexakis

The Thesis of Priyanshu Sharma is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my heartfelt gratitude to my academic advisor, Professor Yue Dong, for their continuous support and guidance throughout my thesis journey. Your profound insights and thoughtful feedback have been crucial to my research. I am immensely grateful for your encouragement and for always being available to discuss ideas and provide solutions. Thank you for believing in my potential and for helping me achieve my goals.

To my family for all the support.

ABSTRACT OF THE THESIS

State of Title IX: A Knowledge Base for Title IX Documentation

by

Priyanshu Sharma

Master of Science, Graduate Program in Computer Science
University of California, Riverside, June 2024
Dr. Yue Dong, Chairperson

Knowledge bases play a vital role in the modern world, offering a systematic and structured approach to integrate various entities, concepts, rules, and relationships associated with real-world information. In this study, we aggregated the official Title IX documentation from 163 institutions, including 13 federal departments, 48 states, and 102 universities. This aggregated data forms the Title IX Knowledge Base (KB), a resource aimed at enhancing understanding and awareness of Title IX. Our approach involves analyzing topics at different levels and extracting fine-grained rules associated with them using Multi-LLM Paradigm.

We evaluated the effectiveness of our Title IX KB using a Retrieval-Augmented Generation (RAG) system, demonstrating superior performance in factuality, semantic similarity, correctness, and query-response relevance compared to naive method. These

results underscore the utility of the Title IX KB in providing access to legitimate, detailed information, facilitating better policy understanding and alignment.

Moreover, this research highlights the potential use of the Title IX KB in training safe and harmless AI assistants under the Constitutional AI Framework. By fostering awareness and enhancing compliance with Title IX regulations across various institutions, the Title IX KB proves to be a valuable tool in promoting informed and equitable practices.

Contents

List of Figures

List of Tables

1. Introduction

In recent years, the discourse surrounding Title IX [17], a landmark federal civil rights law in the United States, has evolved to encompass a myriad of topics ranging from sex-based discrimination to workplace policies. As educational institutions, federal departments, and states navigate the complexities of Title IX compliance, there arises a critical need for comprehensive knowledge bases to consolidate, analyze, and interpret the vast array of associated documentation.

This thesis addresses this need by presenting a systematic approach to constructing a Title IX Knowledge Base (KB). Leveraging official Title IX documentation from a diverse range of entities, including federal departments, states, and universities, this KB serves as a centralized repository of rules, regulations, and policies pertaining to Title IX [18]. Through meticulous data aggregation, preprocessing, and domain analysis, this study lays the foundation for a robust KB that not only aggregates information but also provides insights into core topics and their interrelationships across different entity levels.

Furthermore, this thesis explores the extraction of fine-grained rules from the KB, employing advanced techniques to enhance understanding and alignment with Title IX principles. By evaluating the effectiveness of the KB through a series of experiments and analyses, this research aims to demonstrate its utility in facilitating policy understanding, compliance, and alignment with Title IX regulations. This KB also acts as a unified way of integrating all Title IX rules.

Our work presents a Knowledge Base [20, 21, 22] of Title IX Documentation of around 163 institutions ranging from federal, state to university level. Title IX is a pivotal piece of federal laws with that has profoundly influenced the educational landscape by ensuring gender equality, enhancing opportunities for women in athletics and addressing the issue of school harassment, combating sexual harassment, and fostering inclusive environments. This knowledge base would serve as a centralized repository of information, guidelines, and resources related to Title IX. With this one can extract out the rules and regulations associated with sex-based discrimination, workplace discrimination, etc and can be used as a part of constitution in aligning LLMs for safe human use.

Ultimately, this thesis contributes to the broader discourse on Title IX by providing a systematic framework for knowledge aggregation and interpretation, thereby empowering stakeholders with the information necessary to navigate the intricate landscape of Title IX compliance and policy implementation.

2. Background

There have been numerous efforts to integrate data from various sources to construct Knowledge Bases [1, 2, 3], offering numerous benefits across domains, including improved access to information and consistency. Some of the recent methods include CulturalBank [4], which aims to compile diverse cultural knowledge, and CookingSense [5], which integrates culinary knowledge to enhance RAG Models' [6, 49] performance in culinary tasks.

In a similar way, we utilized the official Title IX documentation from over 150 institutions across three levels: Federal, State, and University. We are pioneers in creating a Knowledge Base in the Title IX domain. Title IX KBs play a crucial role in consolidating rules and regulations concerning sex-based and workplace discrimination, civil rights, and more. Serving as a centralized repository for all Title IX-related rules, its utilization can significantly enhance awareness and understanding of human rights.

Recent work has also focused on the intersection of AI and law. This includes predicting the outcomes of legal cases using the PILOT framework [7]. Additionally, multiple efforts have been made to summarize legal judgments in low-resource environments [8]. Other notable work includes the creation of Lawbench [9], an evaluation benchmark designed to provide a precise assessment of language models in the legal domain. This facilitates the development of more reliable and effective AI tools for the legal industry.

Recently, significant strides have been made in training relatively safe and harmless AI assistants with the assistance of Constitutional AI (CAI) [10, 11, 12]. This approach utilizes a rules-based constitution to align LLMs with no human involvement. Title IX Knowledge Bases can serve as a foundational resource for constructing such a constitution, addressing issues related to discrimination, civil rights, biases, and more. This framework of constitutional AI necessitates far less human interaction compared to other RLHF-based methods [13, 14, 15, 16].

Overall, Title IX Knowledge Bases have a wide range of applications, from training harmless AI to better policymaking. With these objectives in mind, we crawled data from 845 official Title IX documentations [17] to aggregate Title IX-related data into a common knowledge base as part of this thesis. Here are some of the main contributions of this thesis:-

- Aggregated Title IX-related data into a Knowledge Base from the official websites of 163 different institutions, covering 13 Federal Departments [19], 48 States, and 102 Universities.

- Identified, compared, and analyzed the core topics associated with these Title IX documentations at all three different levels.

- Assessed the effectiveness of our Title IX KB using the RAG-based paradigm. Our Rule KB outperformed other methods in terms of semantic similarity, correctness, and query-response relevance.

3. Method

In this section, we will discuss the comprehensive approach to constructing the Title IX Knowledge Base. A Knowledge Base [20, 21, 22] typically serves as a central repository, offering a unified method to understand, generate, and manipulate natural language. Additionally, it consolidates information from multiple sources. For our Title IX Knowledge Base, we compiled rules and regulations from the official websites of over 150+ institutions. By extracting data from these official sources, we ensured the legitimacy and quality of the aggregated Title IX rules.

3.1 Title IX

Title IX [17] is a federal law that prohibits discrimination based on sex, religion, caste, and other characteristics, particularly in educational programs and activities. It provides guidelines for educational institutions to enforce compliance related to Civil Rights [18] and promotes gender equality. Additionally, it offers guidelines for addressing sexual harassment and assault-related issues. Its scope extends from workplace compliance to the responsibilities of Title IX coordinators, addressing issues related to sports, admissions, and recruitment, ensuring equal opportunities for everyone regardless of gender, religion, or caste.

The Title IX Knowledge Base (KB) can help address many related issues by enhancing awareness and understanding of Title IX laws [17, 18]. It serves as a comprehensive resource for educational institutions, legal professionals, and individuals seeking to navigate the complexities of Title IX regulations. By providing clear, accessible information, it supports compliance efforts, promotes gender equity, and aids in the prevention of discrimination and harassment. Furthermore, it offers valuable insights into best practices and emerging trends in Title IX implementation, fostering a culture of continuous improvement and accountability.

3.2 Entity

Entity Type	Examples
Federal	Treasury, Agriculture, etc
State	California, Florida, etc
University	University of Alabama, University of California, Riverside, etc

Table 3.1: List of all the different Entity Type and their examples.

However, before moving forward, we will first discuss the notion of an "Entity" as it is fundamental to understanding the construction of the Title IX Knowledge Base. In this context, the term "Entity" can be defined as any institution at the federal [19], state, or university level that maintains the Official Title IX Documentation. This idea can be further extended to integrate additional possible sources of Title IX Rules at any level. In essence, within our context, an entity can encompass any institution at the federal, state, or university level. Here are some examples of entity levels and their corresponding examples.

Overall, the abstraction of the Entity in our framework is fundamental and allows for the integration of additional Title IX policies and rules from various sources. In the following sections, we will offer a detailed description of the process for constructing the Title IX Knowledge Base.

3.3 Title IX Knowledge Base Construction

Our approach to constructing the end-to-end Title IX Knowledge Base (KB) is comprehensive and methodical. The entire process of building the knowledge base is divided into three main steps: -

1. Data Aggregation
2. Topic Analysis
3. Fine-grained Rules Extraction

Initially, we gathered data from official Title IX websites at various levels. This was followed by topic analysis, a pivotal step in gaining a deep understanding of the aggregated

Title IX Rules. During this phase, we first filtered out noisy data in the preprocessing step and then eliminated unrelated data after identifying clusters of unrelated topics, a common issue in web crawling. The filtered dataset encompasses Title IX implementation at the federal, state, and university levels.

Subsequently, we extracted out rules associated with the previously identified topics using a Multi-LLMs Pipeline. Additionally, we explored and analyzed the performance of our Knowledge Base using RAG settings on the QA Task, and then evaluated their performance using four different metrics: Semantic Similarity, Factuality, Correctness, and Query-Response Relevance.

3.3.1 Data Aggregation

This marks the initial step in constructing our Title IX Knowledge Base. This phase begins with identifying potential sources of Title IX Documentation, followed by data extraction from these sources, and concludes with a preprocessing step. In our case, we identified approximately 845 potential hyperlinks from official websites at the Federal, State, and University levels. This stage primarily consists of two main components: Crawling and Preprocessing. First, let's examine the 'Crawling Data' stage, followed by a discussion of the second stage, 'Preprocessing Data'.

1. Crawling

This step is crucial for the construction of our Title IX KB as it helps us identify potential sources of Title IX Documentation. We primarily focused on crawling the official websites of institutions due to the higher quality and legitimacy of the data. The primary objective of this step is to crawl text data related to Title IX's Rules and Regulations from official websites of Federal, State, and University Level institutions.

Crawled Dataset Information			
Nodes	**Federal**	**State**	**University**
Number of Entity Covered	13	48	102
Average Node Length	1909.81	1026.12	800.64
Crawled Links	56	156	633
Crawl Nodes	6488	20315	43203

Table 3.2: Distribution of Crawled Dataset Across Different Entities.

In total, we have crawled around 845 official websites covering 163 different entities, including 13 Federal Departments [19], 48 State Departments (excluding Wyoming and Kentucky), and 102 Universities. Table 1 shows the distribution of entity in the crawled dataset, preprocessed dataset, and filtered dataset. This includes the number of links crawled, number of text nodes after the crawling stage.

We cover nearly all 13 major US Federal Departments as per the official White House website, 48 states, excluding only Wyoming and Kentucky, which failed due to a Network Exception.

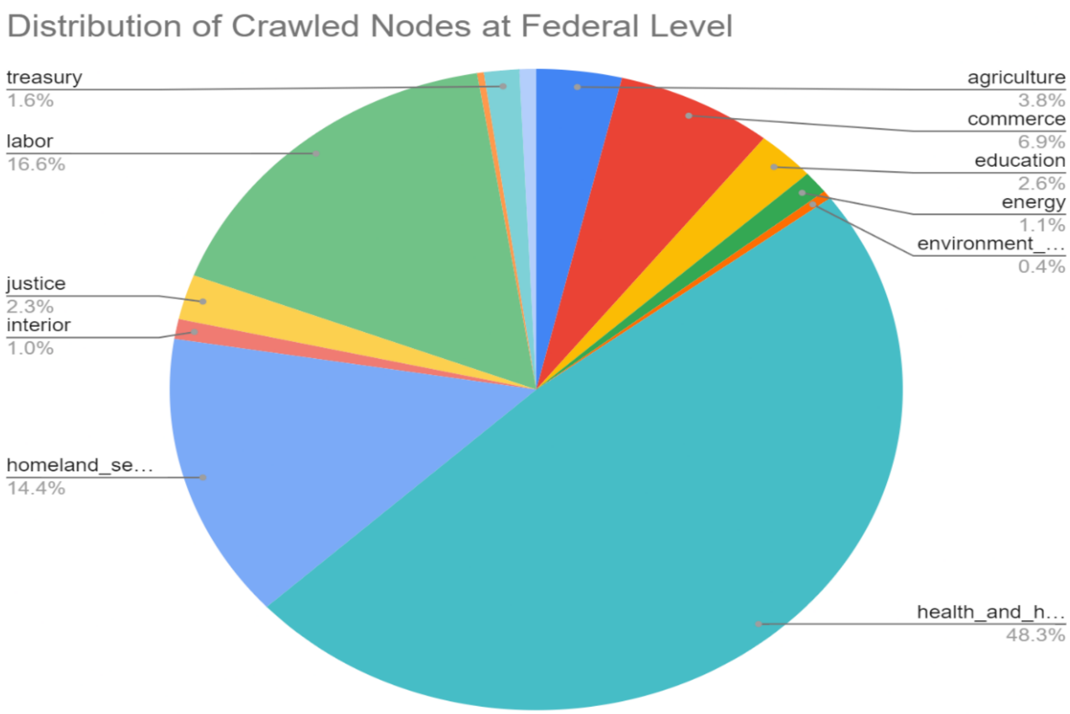


Figure 3.1: Distribution of Crawled data at Federal Level.

Additionally, we include approximately 102 universities, averaging around 2 from each state. Almost all the 845 websites are official institutional websites, with only a few exceptions. The entire crawling process is executed in the Breadth-First Search pattern. Initially, we visit the official Title IX websites of each entity, followed by recursively visiting all Level-1 child nodes of these official websites (hyperlinks present on the official website). To prevent the crawler from revisiting the same link, we maintain a separate counter. This strategy effectively reduces redundant or duplicate crawled data.

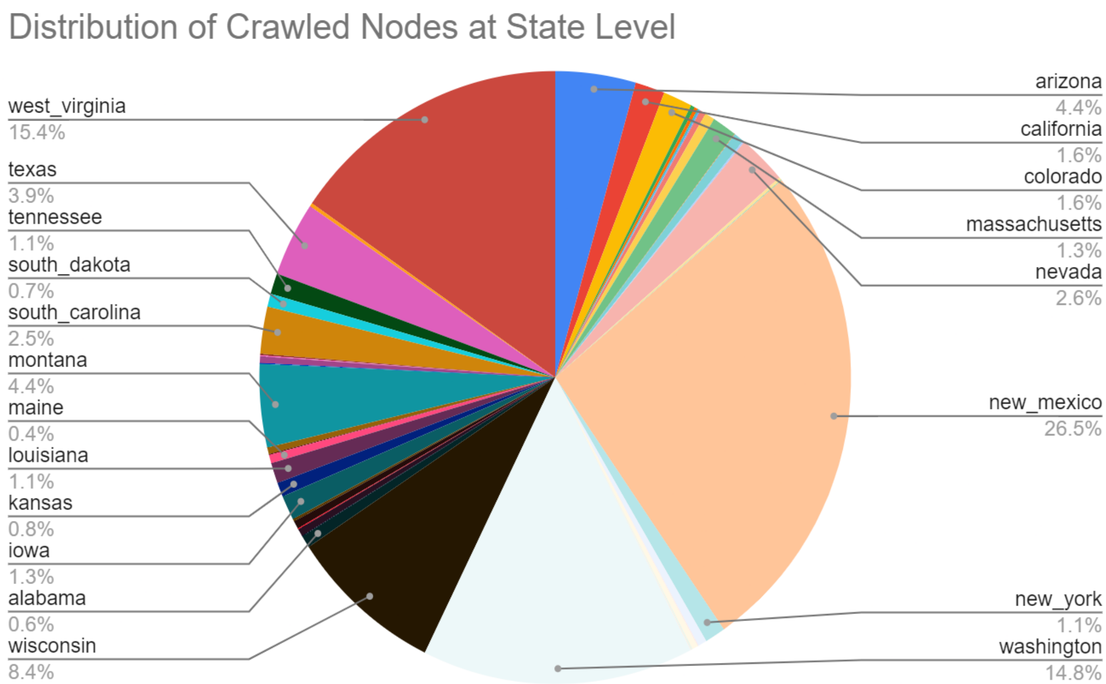


Figure 3.2: Distribution of Crawled Data at State Level.

Figures 3.1, 3.2, and 3.3 depict the entity-wise distribution of the crawled data at the Federal, State, and University levels, respectively. In total, our crawled dataset comprises 70006 text chunks, each sized 1024, extracted from 163 different entities. Specifically, 6488, 20315, and 43203 text chunks are extracted at the Federal, State, and University levels, respectively.

These crawled nodes represent the text dump of visited websites, potentially containing noisy, irrelevant, or corrupt data. And removal of this corrupt data will improve the quality of dataset.

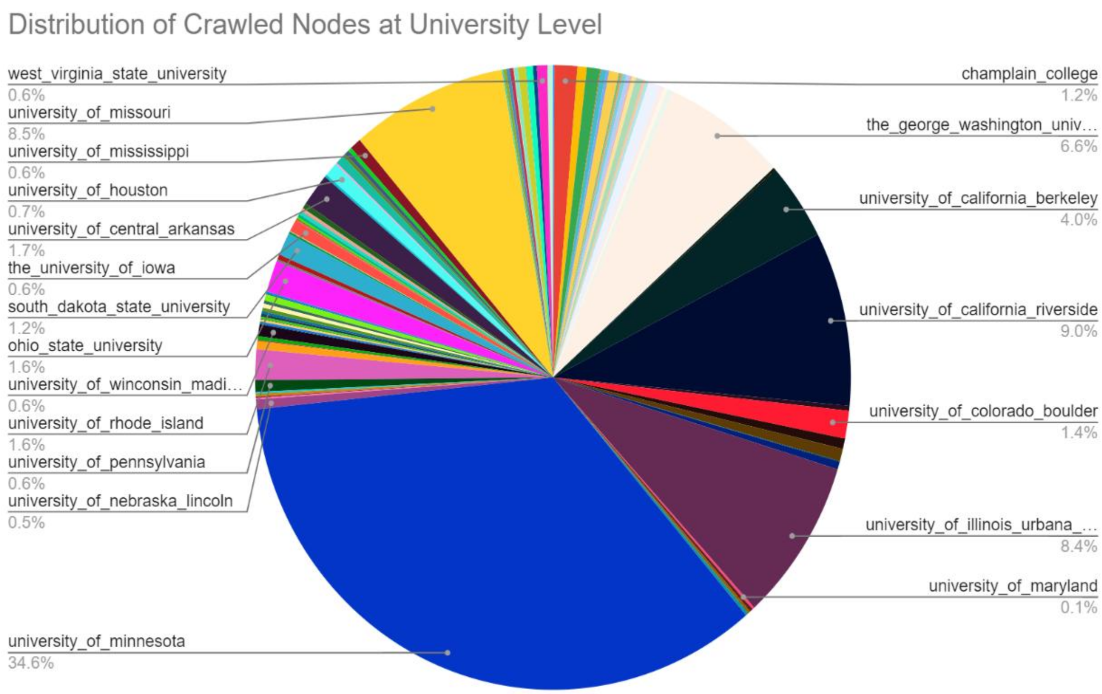


Figure 3.3: Distribution of Crawled Data at University Level.

Therefore, preprocessing these text chunks is crucial to obtain a more accurate understanding of the data. Preprocessing occurs in the second stage of Data Aggregation after collecting the data from the data source, i.e., Official Title IX Websites in our case.

2. Preprocessing

After aggregating data from different sources, such as the Official Title IX Websites in our case, the immediate subsequent step in this process is to preprocess the crawled data. Since we have scraped the website for the data source, which can result in the extraction of noisy data and text data incompatible with UTF encoding, it is common practice to filter out text chunks to remove any noise, corrupt, or null values.

Preprocessed Dataset Information			
Nodes	**Federal**	**State**	**University**
Crawl Nodes	6488	20315	43203
Processed Nodes	6487	19218	43074
Corrupt Nodes	1	1097	129
Corrupt Node %	0.02%	5.30%	0.29%

Table 3.3: Distribution of Preprocessed Dataset Across Different Entities.

The primary contribution of this step lies in refining the crawled data to enhance data quality. This stage also includes formatting the crawled data to remove any HTML tags and to remove URLs present in crawled data. The input for this preprocessing step comprises the crawled text chunks, and the output consists of preprocessed text chunks of improved quality.

As we can see from Figure 3.4, some entities are quite large, such as "University of Minnesota" and "New Mexico," contributing around 14,943 and 5,378 text chunks, respectively, while others are much smaller, like "Michigan" and "Alaska" (Figure 3.5), contributing around 4 text chunks each in the preprocessed dataset.

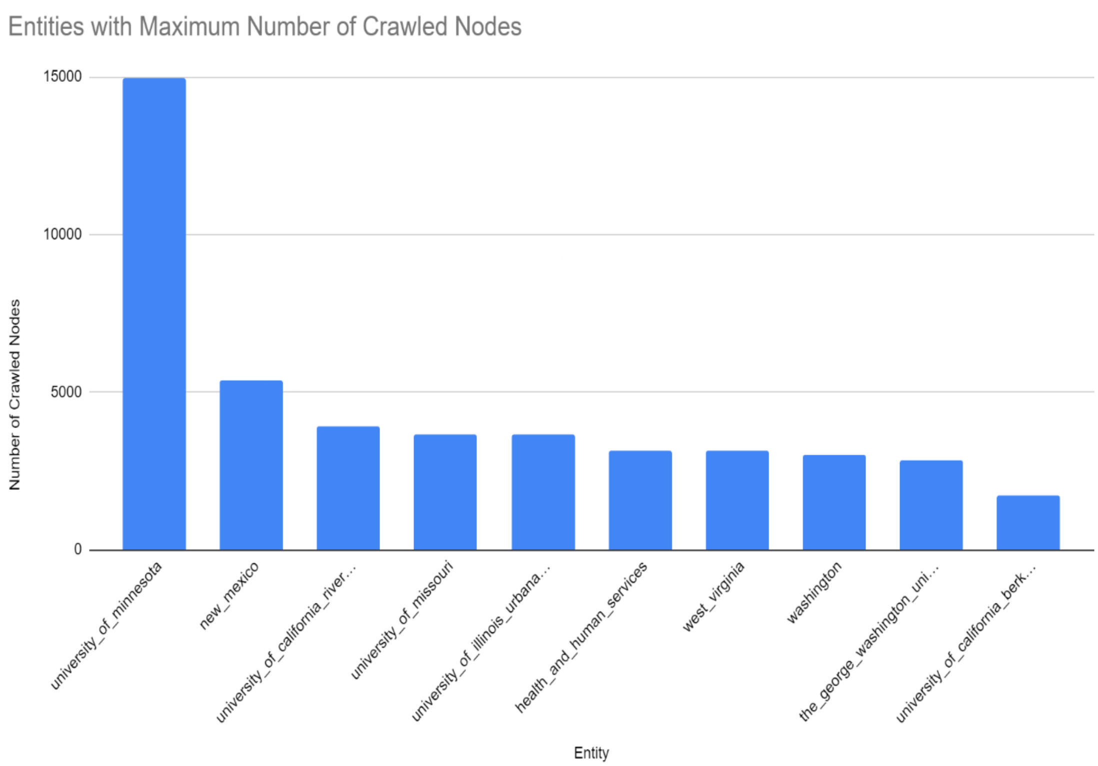


Figure 3.4: Top 10 Entities with Maximum Number of Crawled Nodes.

Overall, the average number of preprocessed text chunks per entity is approximately 429.48, with around 9 entities having fewer than 10 preprocessed text chunks. In total, we have preprocessed around 68,779 text chunks of size 1024, of which 9.43%, 27.94%, and

62.62% comprise Federal, State, and University-level data, respectively. This concludes the Data Aggregation Step, and we will now proceed to the second step of constructing the Title IX KB, namely Topic Analysis of Preprocessed Data. This step plays a crucial role in providing insights into the preprocessed data.

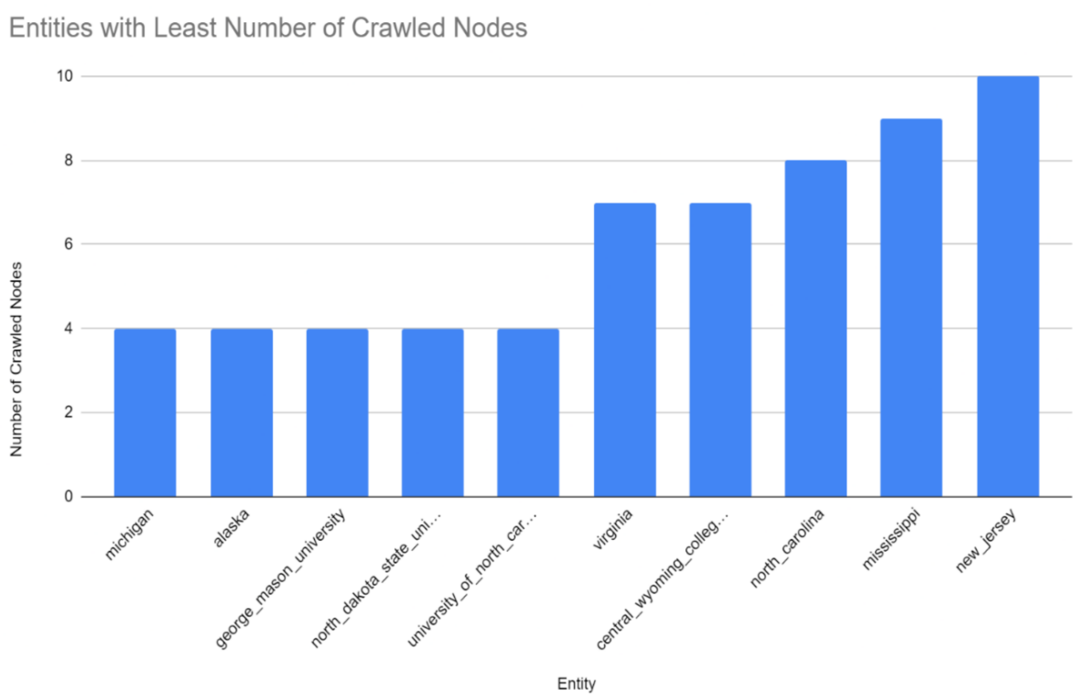


Figure 3.5: Bottom 10 Entities with Least Number of Crawled Nodes.

Next step of Topic Analysis is quite crucial in our end-to-end Title IX KB Construction as it help us in identifying the main topics associated with Title IX.

3.3.2 Topic Analysis

This is the second step in our Title IX KB construction, aiding in the identification of the core domains/areas covered in the Title IX Documentation of different entities. It also indicates the domain in which our extracted rules will reside. It requires the output of the previous step, preprocessed text chunks as input and produces filtered text chunks along with their Cluster information and the associated keywords. The main objective of this step is to identify key domains/subjects covered in the dataset at various entity levels. Later, we will explore how one can use the same common domain to extract the fine-grained Title IX rules associated with them. It is similar to topic modeling [23, 24, 28], which is used to identify abstract topics or themes within a collection of documents. Overall, this step consists of four different stages: Embedding Generation, Dimensionality Reduction, Clustering, and finally Topic Identification.

This entire step begins with ingesting preprocessed text chunks to generate embeddings for each chunk. These embeddings are responsible for clustering semantically similar chunks together. Next, we utilize these extracted embeddings to perform dimensionality reduction and identify the principal components of the dataset.

Dimensionality Reduction plays a pivotal role in our framework. It not only helps us identify the core areas covered by our dataset at all three entity levels but also acts as an additional layer of filtration by removing lower levels of components extracted during the Dimensionality Reduction process. Following this, we perform clustering along these principal components and then proceed with manual Topic Identification using extracted keywords based on TF-IDF Index for each cluster at three different entity levels: Federal, State, and University. Let's now delve into each stage of Core Domain Analysis in depth. We will first identify the Core Domain, after which we will conduct the analysis.

1. Embedding Generation

Embedding Information	
Organization	Jina AI
Model	jina-embeddings-v2-base-en
Architecture	ALiBi bidirectionally + BERT
Dataset Used	Colossal, Cleaned, Common Crawl (C4) by Allen AI
Parameter	137 million

Table 3.4: Embedding Related Information.

To generate the embedding [29], we utilize the processed text chunks and convert them into dense yet low-dimensionality vectors using language models. These vectors are then mapped into a continuous vector space.

This process is primarily responsible for mapping vectors with relatively similar semantic meaning together in the vector space. We employ the "jina-embeddings-v2-base-en" [25] model to extract the embedding of the preprocessed text chunks. This model is based on the BERT [26] architecture and comprises approximately 137 million parameters. It is primarily trained on Allen AI's C4 dataset [27], renowned for its support of longer sequences.

2. Dimensionality Reduction

UMAP Dimensionality Reduction Information	
Hyperparameter	**Values**
Number of neighbors	15
Minimum Distance	0
Number of Components	10
Distance Measure	Euclidean

Table 3.5: Hyperparameter Information for UMAP Dimensionality Reduction.

Previously, we took preprocessed chunks as input and generated the embedding of the corresponding chunks. Generally, these embeddings are present in a high-dimensional space, making it difficult to find hidden patterns within the dataset. We use UMAP [30] for dimensionality reduction, converting higher-dimensional embeddings to a lower-dimensional space. UMAP, a Topological Data Analysis method, excels in capturing complex relationships in non-linear data. It's computationally efficient compared to t-SNE [31] and yields finer results than PCA and SVD, particularly evident in our clustering step. Moreover, UMAP preserves both local and global data structures, controlled by key hyperparameters: number of neighbors and minimum distance.

The Dimensionality Reduction step is performed at different entity levels, namely Federal, State, and Institutional Levels, separately. The hyperparameters used in this process are mentioned in the above table. We use a relatively smaller number of neighbors, specifically 15, to focus more on the local structure. The minimum number of components is around 10 in the hyperparameter, as we expect to obtain 10 core areas at each entity level associated with the text chunks. We also employ the "Euclidean" distance metric to control the distance measure parameter, setting it to 0, as we aim to pack or cluster these core components closely to obtain a more granular detail. Another advantage of this stage is that it helps filter out irrelevant information, thus improving generalization. With this we move to our next stage in Topic Analysis, which is Clustering.

3. Clustering

Clustering the lower-dimensional vectors obtained from Dimensionality Reduction is a crucial step in our process. It helps group semantically similar vectors together to form larger clusters, aiding in uncovering hidden patterns within our dataset. We employ HDBSCAN [32] for its ability to handle different data densities and noise robustness, combining Hierarchical Cluster and DBSCAN [33] methods to capture complex relationships.

HDBSCAN Clustering Information	
Hyperparameter	**Values**
Minimum Cluster Size	5
Distance Measure	Euclidean

Table 3.6: Hyperparameter Information for Clustering.

Setting a minimum cluster size of 5 ensures finer granularity of information. This granular approach allows for deeper insights into related topics within the dataset. The overall picture for our dataset is that we now have clustering information for each preprocessed text chunk.

Another benefit of clustering is that it segregates irrelevant and unrelated information into separate clusters and thus acts as a second layer of filtering. With this, one can quickly identify the topics associated with it by examining a few candidates of a particular cluster and associated keywords.

Following clustering, we proceed to Topic Identification [34], the final stage of Topic Analysis, where we first identify and then analyze and compare common topics at different entity levels.

4. Topic Identification

Topic Identification Information	
Hyperparameter	**Values**
N-Gram Range	1 - 5
English Stop Words	Removed
Frequent Occurring Words	Removed
Top Keywords	30
Diversity Factor	0.3

Table 3.7: Hyperparameter Information for Topic Identification.

In this stage, given the cluster and the associated candidate chunks, we need to identify the keywords associated with them. This stage is fundamental in interpreting the key topics and understanding the relationship between them based on the clusters extracted from the previous stage.

To identify the topics, we first concatenate all the text chunks belonging to a particular cluster. Additionally, we remove English stop words and frequent occurring words before this process. We also consider the possibility of n-grams ranging from 1 to 4 for cluster-wise keyword extraction.

Post this, we utilized the class-based version of the TF-IDF representation [34, 35], which instead of taking the inverse document frequency into account, considers the inverse class frequency. This measures the importance of terms with respect to the entire class, which in our case is the cluster. The output of this stage will assign a few keywords for each cluster based on TF-IDF representation. We mainly focused on the top 30 keywords from each cluster with a diversity factor of 0.3 to identify the important topics at different entity levels. Finally, we manually identify the topics based on the keywords and candidates text chunks for a given cluster.

5. Entity Wise Topic Analysis

In this section, we will primarily compare and analyze the Topic Analysis results at each entity level, namely Federal, State, and Institutional Levels. The main objective behind this step is to identify the core areas/domains covered in the Title IX Implementation at all three different entity levels.

This is an important step before extracting the Fine-grained Rules associated with Title IX Documentation, as it provides us with insight into the domains of the Rules that we are going to extract in the next step.

Topics / Levels	Federal	State	University
Civil Rights (OCR)	Yes	Yes	No
Sexual Harassment Guideline	Yes	Yes	Yes
Federal Enforcement	Yes	Yes	No
Title IX Compliance	Yes	Yes	Yes
Cross Examination	Yes	Yes	Yes
Anti-Discrimination Policy	Yes	Yes	Yes
Workplace Discrimination	Yes	No	Yes
Title IX Coordinator	Yes	Yes	Yes
School Harassment	No	Yes	Yes
Title IX Training	Yes	Yes	Yes

Table 3.8: List of common topics at Federal, State and University levels.

We adhere to the exact same hyperparameter settings for each stage in the Topic Analysis Step across all three different entity levels. Beginning with embedding, followed by dimensionality reduction, clustering, and ultimately Topic Identification, we maintain uniformity throughout all levels.

This final step yields keywords for each cluster, and we manually utilize these keywords to assign topics. For this process, we simply employ the keywords associated with the larger clusters to assign them a common topic.

Federal – At the federal level, most of the Title IX-related rules are extensively covered. A significant portion of data at this level focuses on “Civil Rights”, “Fair Employment Policies”, “Title IX Compliance”, “Joint Complaints”, “Grievance Process”, “Title IX Coordinator”, etc. It comprehensively addresses the topics of Federal Regulations such as CFR (Code for Federal Regulations) and CRCL (Civil Rights and Civil Liberties). As an output of this step, we identify around 180 different topics covered in the Federal Level Title IX Implementation.

State – At this level, the predominant topics include “Pregnancy Related”, “Sports Related Discrimination”, “Cross Examination”, “Party Witness Compliances”, “Hearing Process”, “Sexual Harassment Prevention Training”, etc. Additionally, we observe that these rules are available in multiple languages such as Bengali, Russian, Italian, etc., which was not the case with the Federal Level Data. Overall, for the state level, we identify around 303 different clusters, each with different topics. The Cluster Distribution for the State Level differs significantly from the Federal Level, with relatively smaller cluster sizes compared to the Federal one.

University – At this level, approximately 385 different topics are generated, with most of them related to “Title IX Compliance”, “Formal Written Complaint”, “Appeal Process” “Equal Employment Opportunity”, “Sexual Harassment”, Reporting”, “Workplace Discrimination”, “Sexual Harassment Awareness Training”, etc. It also covers a wide range of domains, from sports to the admission process.

6. Similarity vs Difference

Most of the topics are common across different entity levels. These include “Title IX Compliance”, “Sexual Harassment Policy”, “Workplace Compliance”, “Title IX Grievance Process”, etc. Although there are similarities, there are also differences. For example, Federal Level Data focuses more on topics like “Civil Rights”, “Enforcement covered by federal agencies such as the Office of Civil Rights”, “Federal Assistance covered under Title IX”, “Pregnancy-related guidelines”, etc. State Level Data covers a range of topics like “Sport-Related Discrimination”, “Workplace Policies”, “Guidelines for genetic monitoring”, “Sex-based discrimination”, etc. One of the most significant differences between federal and state-level data is the availability of Title IX guidelines in multiple languages, including Italian, Russian, Bengali, etc. University-level data primarily focuses on “Title IX Policies”, “Sexual Harassment Training Process”, “Title IX Coordinator responsibilities”, “Athletic Enrollment and Scholarships Guidelines”, etc.

7. Filtered Dataset

Clustering the preprocessed nodes helps us identify the topics associated with the Title IX documentation at different entity levels. As a result of this clustering process, some of the clusters are entirely composed of irrelevant, and of noisy data. This also assists us in providing another level of filtering. For this, we manually identify the noisy/irrelevant clusters with the help of topics, keywords, and candidate documents identified during the clustering process. We use this manual process to gain a more detailed understanding of the cluster composition. Some of the cluster topics are mentioned in Table 3.8.

Dataset Information			
Nodes	**Federal**	**State**	**University**
Crawl Nodes	6488	20315	43203
Processed Nodes	6487	19218	43074
Filtered Nodes	5820	9724	14746

Table 3.9: Number of filtered nodes present at different entity levels.

As a result of this manual process, we filtered out approximately 5820, 9724, and 14746 nodes at the Federal, State, and University levels respectively. With this, it marks the end of the filtering process and the topics analysis step. Our final dataset comprises around 30290 nodes, which is approximately 44% of the original dataset. With the collection of the filtered dataset, we have completed the basic construction of the Title IX KB. Now, we will utilize the On-Demand Information Extraction pipeline to extract the fine-grained details associated with our dataset.

3.3.3 Fine-grained Rules Extraction

The primary objective of this step is to obtain a more granular level of rules related to Title IX Documentation. We utilize the Language model to extract fine-grained rules associated with each filtered node. To accomplish this, we extract 40 different entities from the filtered dataset for this process. Among these 40 entities, we consider 10 from the federal level, 16 from states, and 14 from the university level.

Recently, we have witnessed the Instruct and Extract Technique [36], which is based on the “On-Demand Information Extraction” paradigm, assisting us in extracting information from the provided text chunks in the table. We follow a similar pattern with a modified prompt specifically for rules extraction. The output of this stage is a list of rules for each filtered text chunk. Each rule consists of a “Rule Title” along with information about the topics in the filtered nodes.

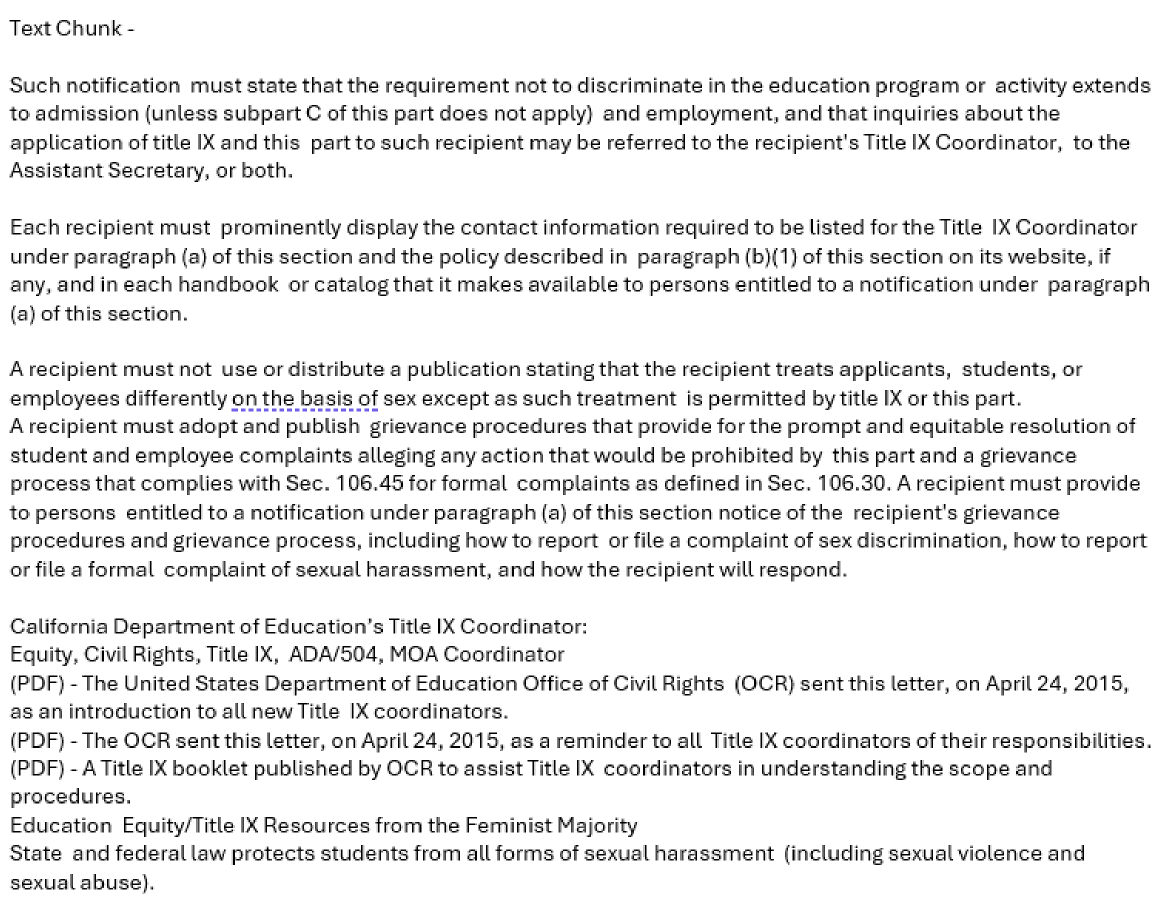
For this process, we employ the Multi-LLM [37, 38, 39] paradigm, which comprises two layers of rules extraction, each with a different model. The first level of rules extraction is based on basic prompting.

```
"""
You are a helpful assistant, who help in data extraction. Here is the example query -

Example -

Given the context information and not prior knowledge, answer the query.

Context: "Strawberries are a popular fruit known for their vibrant red color and sweet, juicy flavor. One of the most widely consumed fruits,
apples come in various colors, including red. They are round in shape and crisp, and sweet in taste. They are highly rich in Dietary fiber and
Vitamin C. Strawberries are also quite rich in Vitamin C and have wonderful antioxidants properties. Bananas are elongated, where as
Strawberries are Heartshaped, slightly curved fruits that have a thick, protective peel and soft, sweet flesh. They are quite rich in Potassium and in
Vitamin B6."

Instruction: Extract out the detailed information from the given context in the Table.

Extracted Table:
------------------------------------------------------------------------------------
| Keyword     | Information                                              |
------------------------------------------------------------------------------------
| Strawberries | It is popular fruit and vibrant red in color.            |
|              | Sweet, juicy in flavor.                                  |
|              | Rich in Vitamin C and have antioxidants properties.      |
|              | Heartshaped.                                             |
| Apples       | Comes in various colors, including red.                  |
|              | Round in shape and crisp, sweet in taste.                |
|              | Highly rich in Dietary fiber, vitamin C.                 |
| Bananas      | Bananas are elongated, slightly curved in shape.         |
|              | They have thick and protective peel and soft, sweet flesh.|
|              | Rich in Potassium, vitamin B6.                           |
------------------------------------------------------------------------------------

Now, Given the context information and not prior knowledge, answer the following query.

Context: "{context_str}"

Instruction: Extract out the detailed information from the given context in the Table.

Extracted Table:
```

Figure 3.6: Prompt used in Basic Rules Extraction Process.

While the second level of extraction is based on refined prompt, which incorporates the response of the first level of extraction along with the filtered text chunks. These two levels of extraction aid us in identifying even the smallest details as well as complete information about the mentioned rules in the filtered nodes. The idea is to squeeze out the complete information from the filtered nodes.

1. Basic Rules Extraction



Text Chunk -

Such notification must state that the requirement not to discriminate in the education program or activity extends to admission (unless subpart C of this part does not apply) and employment, and that inquiries about the application of title IX and this part to such recipient may be referred to the recipient's Title IX Coordinator, to the Assistant Secretary, or both.

Each recipient must prominently display the contact information required to be listed for the Title IX Coordinator under paragraph (a) of this section and the policy described in paragraph (b)(1) of this section on its website, if any, and in each handbook or catalog that it makes available to persons entitled to a notification under paragraph (a) of this section.

A recipient must not use or distribute a publication stating that the recipient treats applicants, students, or employees differently on the basis of sex except as such treatment is permitted by title IX or this part.
A recipient must adopt and publish grievance procedures that provide for the prompt and equitable resolution of student and employee complaints alleging any action that would be prohibited by this part and a grievance process that complies with Sec. 106.45 for formal complaints as defined in Sec. 106.30. A recipient must provide to persons entitled to a notification under paragraph (a) of this section notice of the recipient's grievance procedures and grievance process, including how to report or file a complaint of sex discrimination, how to report or file a formal complaint of sexual harassment, and how the recipient will respond.

California Department of Education's Title IX Coordinator:
Equity, Civil Rights, Title IX, ADA/504, MOA Coordinator
(PDF) - The United States Department of Education Office of Civil Rights (OCR) sent this letter, on April 24, 2015, as an introduction to all new Title IX coordinators.
(PDF) - The OCR sent this letter, on April 24, 2015, as a reminder to all Title IX coordinators of their responsibilities.
(PDF) - A Title IX booklet published by OCR to assist Title IX coordinators in understanding the scope and procedures.
Education Equity/Title IX Resources from the Feminist Majority
State and federal law protects students from all forms of sexual harassment (including sexual violence and sexual abuse).

Figure 3.7: Example of Filtered text chunk.

The focus of basic rules extraction is to identify the key rules mentioned in the given text chunks and extract the information into the table as output. This entails identifying even the smallest details about the given rules from the text chunks. We use the following Prompt (Figure 3.6) to gather this information.

Keyword	Information
Notification	Must state that discrimination is not allowed in education program or activity.
	It should also mention that inquiries can be referred to the Title IX Coordinator or Assistant Secretary.
Title IX Coordinator	Contact information must be displayed prominently on website, handbooks and catalogs.
Recipient	Must not use or distribute publication that states treatment of applicants based on sex.
	Must adopt and publish grievance procedures for resolution of complaints.
	Must provide ways to report or file a complaint of sex discrimination or sexual harassment.
California Department of Education	Title IX Coordinator's designated title and department.
Resources	Links to various PDF documents and websites for further information.

Figure 3.8: Output of Filtered text chunk after Basic Rule Extraction.

With these objectives in mind, we utilize the "Command-R" [40] model by Cohere to conduct the Basic Rules Extraction. It was one of the best models of its time, addressing the challenge of "identifying the needle in a haystack problem". Here are a few examples of the Basic Rules Extraction.

2. Refined Rules Extraction

This stage largely focuses on extracting the remaining rules and refining the rules obtained from the basic extraction. If any rule is missed in the Basic Extraction, then this phase will add the missing one to the extracted rules. Overall, this stage helps us generate a more granular level of rules compared to the previous stage with the help of "Llama2" [41] by Meta.

```
"""
The original query is as follows: "Extract out the detailed information from the given context in the Table."

We have provided an existing table: {existing_answer}
We have the opportunity to refine the existing table (only if needed) with some more context below.
------------
{context_str}
------------
Given the new context, if it has more detailed information
then add the new detailed information in the existing table. \
If the context isn't useful, return the original table.
Refined Table: \
"""
```

Figure 3.9: Prompt used in Refined Rules Extraction Process.

For Refined Rules Extraction, we use a refined prompt that takes the extracted table as input along with the respective text chunks and adds more/missing rules to the table based on the text chunks. The number of extracted rules is much higher and more granular compared to the previous stage.

```
---------------------------------------------------------------------------------------------------------------------------------------------------------
| Keyword                          | Information                                                                                                          |
|----------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Notification                     | Must state that discrimination is not allowed in education program or activity, and that inquiries can be referred to
the Title IX Coordinator or Assistant Secretary.                                                                                                          |
| Title IX Coordinator             | Contact information must be displayed prominently on website, handbooks, and catalogs. The California
Department of Education's Title IX Coordinator is Equity, Civil Rights, Title IX, ADA/504, MOA Coordinator. (PDF) - The United States
Department of Education Office of Civil Rights (OCR) sent a letter to all new Title IX coordinators on April 24, 2015, as an introduction. (PDF) -
The OCR sent a reminder to all Title IX coordinators on April 24, 2015, of their responsibilities. (PDF) - A Title IX booklet published by OCR to
assist Title IX coordinators in understanding the scope and procedures.                                                                                  |
| Recipient                        | Must not use or distribute publication stating that the recipient treats applicants, students, or employees differently
on the basis of sex except as such treatment is permitted by title IX or this part.                                                                      |
| Grievance Procedures             | A recipient must adopt and publish grievance procedures that provide for the prompt and equitable resolution of
student and employee complaints alleging any action that would be prohibited by this part, and a grievance process that complies with Sec. 106.45
for formal complaints as defined in Sec. 106.30. A recipient must provide to persons entitled to a notification under paragraph (a) of this section
notice of the recipient's grievance procedures and grievance process, including how to report or file a complaint of sex discrimination, how to
report or file a formal complaint of sexual harassment, and how the recipient will respond.                                                              |
| California Department of Education | The California Department of Education's Title IX Coordinator is Equity, Civil Rights, Title IX, ADA/504,
MOA Coordinator. (PDF) - The United States Department of Education Office of Civil Rights (OCR) sent a letter to all new Title IX coordinators
on April 24, 2015, as an introduction. (PDF) - The OCR sent a reminder to all Title IX coordinators on April 24, 2015, of their responsibilities.
(PDF) - A Title IX booklet published by OCR to assist Title IX coordinators in understanding the scope and procedures.                                   |
| Resources                        | Links to various PDF documents and websites for further information.                                                 |
---------------------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 3.10: Output of Filtered text chunk after Refined Rule Extraction.

Rules Statistics			
	Federal	**State**	**University**
Number of Entity	10	16	14
Extracted Rules	6724	7968	5962
Number of Nodes	1132	1276	704

Table 3.10: Total number of Rules Extracted Refined Rules Extraction .

Here are some of the Rules Statistics for 40 different entities after the Refined Rules Extraction Process. We have also included entity-wise analysis of the extracted rules. With this, we conclude our end-to-end pipeline, which we used to extract the fine-grained rules associated with Title IX Documentation from various entities. This marks the completion of building our Title IX Knowledge Base. Next, we will discuss a few experiments designed to demonstrate the significance of having such a Knowledge Base.

4. Experiments

In this section, we will explore the experiments conducted to demonstrate the usability of our Title IX Knowledge Base. We used our previously sampled data of 40 different entities as the dataset for generating Q/A pairs. This was accomplished using the Phi-3 [42] Model, which randomly selected three consecutive filtered nodes from the 40 sampled entities as context to generate the Q/A pairs. We collected around 347 Q/A pairs for the sampled entities.

Next, we manually identified context-specific Q/A pairs that were highly relevant to the provided contexts. We tagged all the Q/A pairs by topic and retained only the 174 most general Q/A pairs. These include 44, 64, and 66 Q/A pairs from the Federal, State, and University levels, respectively, covering a wide range of topics related to Title IX. We then used these pairs to conduct our experiment, ensuring a comprehensive evaluation of the Title IX Knowledge Base.

We then proceeded with our Q/A experiment, which was conducted using three different configurations. These configurations included: comparison with the basic language model (Gemma-2B) [43], referred to as the Naïve configuration; the RAG System [6, 49] based on the Gemma model using our collected Title IX Knowledge Base (covering data for only the 40 sampled entities), referred to as the Basic KB Configuration; and the

RAG System [6, 49] based on Gemma-2B [43] using the previously extracted fine-grained rules, referred to as the Rules KB configuration. Our first configuration was based solely on the language model without any Title IX Knowledge Base. This evaluation, conducted with the Gemma model, served as a Baseline Evaluation of our experiments, aiming to test the internal knowledge of LLMs about Title IX rules.

	Naive Method	**Basic KB**	**Rules KB**
Language Model	Gemma	Gemma	Gemma
Vector Database	-	FAISS	FAISS
Index	-	Flat Index	Flat Index
Knowledge Base	-	Filtered Text Chunks	Rules
Top-K Retrieval	-	10	10

Table 4.1: Different configurations used for Evaluation.

The second configuration, Basic KB, used the actual filtered text chunks from our Title IX Knowledge Base as a data source. This setup required a RAG system, embedding the filtered text chunks from the 40 sampled entities into a FAISS [44, 45] Vector Database [48, 50] to answer questions. It utilized a Flat Index (L2) based on Euclidean Distance to calculate semantic similarity between the questions and the retrieved text chunks. We retrieved the top 10 most similar nodes for both this configuration and the next one.

Our final configuration, the Rules KB, was also a RAG system [6, 49] but was based entirely on extracted rules. These rules provided a more granular level of information from the given text chunks. This setup used a similar configuration to the second one, with an L2 Index and a FAISS [44, 45] Vector Database [48, 50] for embedding the extracted rules. Again, we retrieved the top 10 most semantically similar nodes based on the given question.

5. Results

Let’s first explore all the different metrics we have used to evaluate the three different configurations, i.e. – Naïve, Basic KB, and Rule KB. We first evaluate theses configuration on Factuality, Semantic Similarity, Correctness and Query-Response Relevance. These all are language model-driven metrices. Post that, we also perform the Hit Rate Analysis for our later two configurations, i.e. – Basic KB, and Rule KB. We will observe that our Rule KB based methods outperforms the other methods in Semantic Similarity, Correctness, Query-Response Relevance and Hit Rate, while Basic KB is better in terms of Factuality. Let’s now discuss all these results in detail.

5.1 Preliminary Analysis

In this section, we will conduct and analyze the results of preliminary analysis on our three different configurations: Naïve, Basic KB, and Rule KB. Here, we measured Factuality using the "FactKB" [46] language model, and Semantic Similarity, Correctness, and Query-Response Relevance using the "Mistral-7b" [47] model. Factuality and Semantic Similarity scores range from 0 to 1, while Correctness and Query-Response Relevance scores range from 0 to 5. We applied the exact same settings as outlined in the experiments section.

Factuality – We assessed factuality on a scale of 0 to 1 using the FactKB [46] model, specifically trained to measure the factual consistency of generated responses. We evaluated the factuality of the generated responses relative to the ground truth obtained during Q/A Generation. Our findings indicate that the Basic KB method outperforms the other methods in terms of factual consistency of generated answers. Specifically, the Naïve method, Basic KB, and Rules KB scored approximately 82%, 92.1%, and 88.6%, respectively.

	Naive	Basic KB	Rules KB
Factuality	0.823	0.921	0.886
Semantic Similarity	0.41	0.46	0.47
Correctness	4.533	4.202	4.6
Query-Response Relevance	4.7	4.724	4.77
Hit Rate	-	0.649	0.701

Table 4.2: Results for Naïve, Basic KB and Rules KB Methods.

Semantic Similarity – In this analysis, we measured the semantic similarity between the embeddings of the generated responses and the ground truth using the Mistral-7B [47] model and the 'jina-embeddings-v2-base-en' [25] embedding. Similarity scores ranged from 0 to 1. Our findings indicate that the Rules KB outperforms the other methods significantly. Specifically, for the Rules KB, semantic similarity was approximately 47%, while for the Naïve and Basic KB methods, it was around 41% and 46%, respectively. Overall, Rules KB is much better in retrieving semantic similar chunks as compared to other methods.

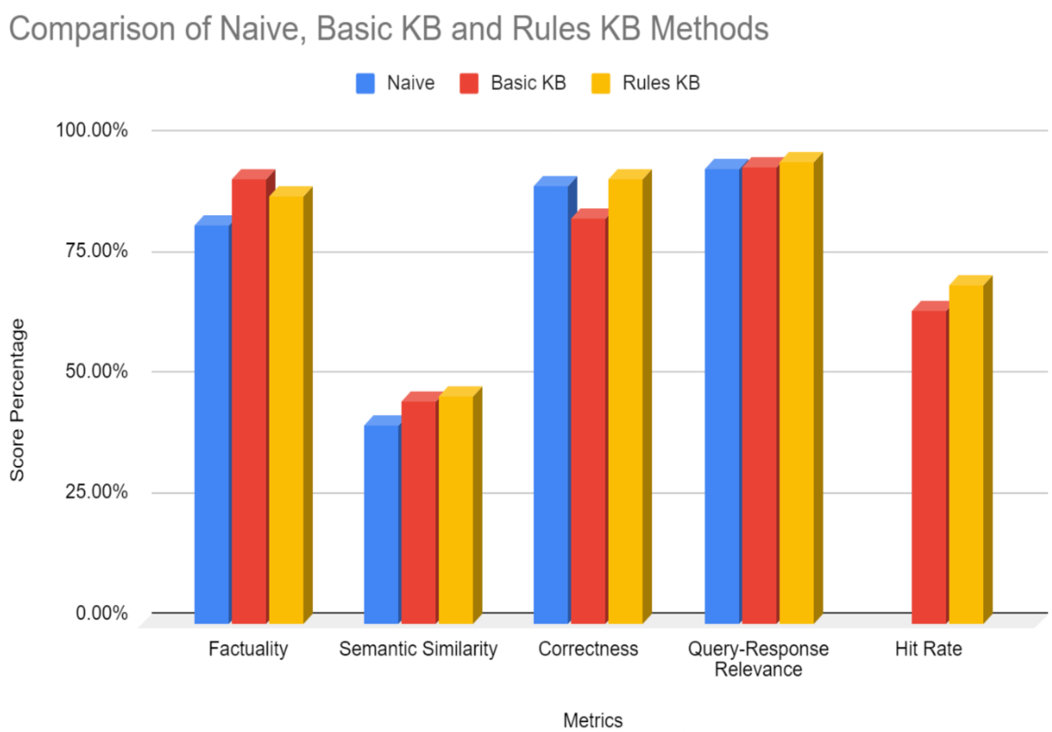


Figure 4.1: Comparison of Performance of Naïve, Basic KB and Rules KB.

Correctness – This metric evaluates the accuracy of the generated responses relative to the given question and ground truth. It assesses whether the responses align with the question and ground truth. Scores ranged from 0 to 5, measured using the Mistral-7B [47] model. In this aspect, the Rules KB method outperforms the Naïve and Basic KB methods significantly. Specifically, the Naïve, Basic KB, and Rules KB methods scored approximately 90.66%, 84.04%, and 92.06%, respectively, in terms of correctness.

Query-Response Relevance – This metric assesses the relevance of the generated response to the given question. Responses that effectively address most of the subject matter associated with the question receive higher scores in this metric. Evaluation was conducted using the Mistral-7B [47] model on a scale of 0 to 5. Once again, the Rules KB method outperforms the other methods. The Rules KB scored approximately 95.40% in terms of query-response relevance, slightly higher than the Naïve and Basic KB methods, which scored around 94.17% and 94.48% respectively.

Hit Rate – We also performed a Hit Rate Analysis for our Basic KB and Rules KB, as they both use our Title IX Knowledge Base as a data source. Out of 174 questions, 113 and 122 were correctly answered by Basic KB and Rules KB, respectively. Consequently, the Hit Rate for Rules KB was better than that for Basic KB. Specifically, the Hit Rate for Rules KB was approximately 70.19%, while for Basic KB, it was around 64.94%. This improvement in Rules KB is primarily due to the better retrieval of more granular rules compared to the entire text chunk. We did not analyze the Hit Rate of the Naïve method, as it does not use any data source for generating answers, relying solely on its prior knowledge.

5.2 Comparative Analysis

In this section, we will compare and analyze the results obtained during our evaluation process. We will delve into the performance of our Basic KB and Rules KB methods to understand the significance of the Title IX Knowledge Base. Additionally, we will examine the role of Fine-grained Rules Extraction in enhancing the performance of our Rules KB in terms of Semantic Similarity, Correctness, and Query-Response Relevance.

	Naive	Basic KB	Rules KB
Factuality	82.30%	92.10%	88.60%
Semantic Similarity	41.00%	46.00%	47.00%
Correctness	90.60%	84.00%	92.00%
Query-Response Relevance	94.10%	94.40%	95.40%
Hit Rate	-	64.90%	70.10%

Table 4.3: Results for Naïve, Basic KB and Rules KB Methods in Percentage.

Let's begin with Factuality. Our Basic KB outperforms both the Rules KB and Naïve Methods in this metric. However, the Rules KB also performs better compared to the Naïve method. The primary reason behind this is that methods like Basic KB and Rules

KB leverage our Title IX Knowledge Base. This enables them to retrieve authentic and relevant information from the Knowledge Base, whereas the Naïve method relies solely on its prior knowledge to answer the given question.

The absence of a Knowledge Base can cause the Naïve Method to hallucinate and generate less relevant information, resulting in a decrease in factuality. In the case of the Rules KB, we observe a slight decrease in performance compared to the Basic KB in terms of Factuality. This could be attributed to the use of fine-grained rules instead of actual filtered text chunks. Fine-grained rules decompose the filtered text chunks into smaller components, requiring the retrieval of more rules to achieve the same level of factuality.

For other metrics such as Semantic Similarity, Correctness, Query-Response Relevance and Hit Rate, our Rules KB outperforms other methods significantly. We suspect two main reasons for this. First, the extraction of fine-grained rules from the filtered dataset helps us decompose the entire text chunks into a set of rules. These rules consist of two main components: a Rule Title separated by a colon and the rule's details. Each rule has its own Rule Title, which plays a major role in retrieving higher-quality related data based on the subject mentioned in the given question.

Additionally, it provides a more granular level of information compared to the Basic KB, which uses filtered text chunks as embeddings rather than fine-grained rules. With the Rules KB, we almost always retrieve much more semantically similar or related information, whereas with the Basic KB, it may happen that the retrieved filtered text chunks contain much less relevant data compared to the chunks retrieved in the Rules KB, yet they are still being retrieved.

Overall, we have determined that methods like Rules KB and Basic KB outperform Naïve methods in terms of performance for the following reasons:-

1. Our methods, such as Rules KB and Basic KB, are supported by our Title IX Knowledge Base, enabling us to retrieve more legitimate information compared to the Naïve Method.

2. The use of fine-grained rules, as opposed to entire filtered text chunks, provides, and embeds a much more granular level of information in vector space. As a result, it assists us in retrieving higher-quality extracted rules.

3. The utilization of Rule Titles in extracted rules aids in pruning our search space by directly retrieving the exact rules associated with the topic mentioned in the query.

6. Discussion

In this thesis, we contributed to the creation of a Knowledge Base, which consists of Title IX documentation [17, 18] from approximately 13 Federal Departments [19], 48 States, and 102 Universities. This Knowledge Base is significant because it can play a vital role in raising awareness and providing a better understanding of Title IX documentation. The Title IX Knowledge Base covers a wide range of topics, including "Title IX Rules," "Anti-Discriminatory Policies," and "Civil Rights." Additionally, this Knowledge Base can be utilized to extract rules for a Constitutional AI (CAI) Framework [10, 11, 12], aligning LLMs on topics such as "Discrimination Policy" and "Civil Rights."

To implement this Title IX Knowledge Base, we began with the Data Aggregation Step, which involved crawling the official Title IX documentation of 163 institutions to form the crawled dataset. After crawling, we preprocessed the entire dataset, removing any empty nodes, hyperlinks, and HTML tags, resulting in our preprocessed dataset. Following this, we conducted domain analysis of the preprocessed dataset, including Dimensionality Reduction [30, 31] and Clustering of the core components [32, 33]. This step helped us identify the core components of the dataset and their clusters at the Federal, State, and University levels. It provided in-depth information about the relationships between different clusters and acted as a filter for removing irrelevant or unrelated data. Additionally, we identified the keywords associated with each cluster using class-based

TF-IDF representation [28, 29], which aided in the manual identification of topics associated with the clusters. Overall, it helps us in identifying common Title IX topics at all three different levels.

We found that certain topics, such as "Title IX Compliance," "Anti-Discriminatory Policy," "Title IX Coordinator's Responsibility," "Sexual Harassment Training," and "Workplace Compliances," were common across all entity levels. However, specific topics varied by entity: federal level topics included Civil Rights, Federal Regulation, and Federal Assistance for Title IX; state-level topics focused on sport-related discrimination. Additionally, Title IX policies were available in multiple languages at the State and University levels, unlike at the Federal level. We also extracted fine-grained rules associated with Title IX using a Multi-LLM Paradigm [37, 38, 39], with two levels of on-demand information extraction. These extracted rules provided in-depth information about the topics related to Title IX regulations.

We experimented with three different configurations to demonstrate the usability of our Title IX Knowledge Base. This Knowledge Base not only allows users to identify the rules and regulations associated with Title IX-related topics but also facilitates the exploration of complex issues such as Civil Rights, the relationship between Title VI and Title IX, pregnancy-related discrimination policies, and gender monitoring policies. While our approach performs well on the Title IX KB, it also has some limitations. Despite the granularity of our rules, they often lack connections with other rules, making it challenging to extract complete sets of rules. Exploring a more promising pipeline for extracting fine-grained rules using a graph-like interface to preserve relationships between different rules could address this issue.

Once we have a more robust pipeline for rule extraction, we can proceed to construct a constitution for LLMs using these rules. This constitution, along with the rules, will aid in aligning LLMs for safe human use and ensuring their harmlessness for day-to-day applications. Additionally, we can utilize these rules to assess the political alignment of various LLMs, determining whether they align with left or right ideologies. Furthermore, exploring agentic evaluation of these Title IX rules can reveal policy differences among federal, state, and university levels, enabling comparison and identification of shortcomings in the policies of specific entities.

7. Conclusion

In conclusion, our study presents a more generalizable and systematic approach for creating the Title IX Knowledge Base. Through our pipeline, we curated a comprehensive Title IX Knowledge Base using official documentation from 163 entities, encompassing 13 federal departments, 48 states, and 102 universities. We identified numerous key topics and their associated rules within Title IX across all three entity levels.

Additionally, we introduced a Multi-LLM Pipeline for extracting fine-grained rules on various topics. Finally, we employed an RAG-based system to evaluate the effectiveness of our Knowledge Base. Our findings indicate that our Title IX Knowledge Base, implemented with the RAG system, outperforms other methods lacking our Title IX Knowledge Base.

Overall, the utilization of our Title IX Knowledge Base along with the Fine-grained Rules Extraction pipeline can significantly enhance the quality of retrieved data and can be used in exploring and understanding the complex rules of Title IX Laws.

Acknowledgements

This work was supported by the NLP Lab of University of California, Riverside. This research uses the resources from MLNLP cluster. I would like to take this opportunity to thank Victor Hill for providing his undivided support in solving the issue related to cluster.

Bibliography

[1] Vuong, T., Hoang, M., et.al., Constructing a Knowledge Graph for Vietnamese Legal Cases with Heterogeneous Graphs. (https://arxiv.org/abs/2309.09069).
[2] Wiratunga, N., Abeyratne, R., et al., CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. (https://arxiv.org/abs/2404.04302).
[3] Yu, Z., Li, Y., An Argumentation-Based Legal Reasoning Approach for DL-Ontology. (https://arxiv.org/abs/2209.03070).
[4] Shi, W., Li, R., et al., CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. (https://arxiv.org/abs/2404.15238).
[5] Choi, D., Gim, M., et al., CookingSense: A Culinary Knowledgebase with Multidisciplinary Assertions. (https://arxiv.org/abs/2405.00523).
[6] Lewis P, Perez, E., et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. (https://arxiv.org/abs/2005.11401).
[7] Cao, L., Wang, Z., et al., PILOT: Legal Case Outcome Prediction with Case Law. (https://arxiv.org/abs/2401.15770).
[8] Liu, S., Cao, J., et al., Low-Resource Court Judgment Summarization for Common Law Systems. (https://arxiv.org/abs/2403.04454).
[9] Fei, Z., Shen, Z., et al., LawBench: Benchmarking Legal Knowledge of Large Language Models. (https://arxiv.org/abs/2309.16289).
[10] Bai, Y., Kadavath, S., et al., Constitutional AI: Harmlessness from AI Feedback. (https://arxiv.org/abs/2212.08073).
[11] Kundu, S., Bai, Y., et al., Specific versus General Principles for Constitutional AI. (https://arxiv.org/abs/2310.13798).
[12] Petridis, S., Wedin, B., et al., ConstitutionalExperts: Training a Mixture of Principle-based Prompts. (https://arxiv.org/abs/2403.04894).
[13] Park, C., Liu, M., et al., RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. (https://arxiv.org/abs/2405.00254).
[14] Dong, H., Xiong, W., et al., RLHF Workflow: From Reward Modeling to Online RLHF. (https://arxiv.org/abs/2405.07863).
[15] Hu, J., Wu, X., et al., OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. (https://arxiv.org/abs/2405.11143).
[16] Xiao, J., Li, Z., et al., On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization. (https://arxiv.org/abs/2405.16455).
[17] Title IX and Sex Discrimination, US Department of Education. (https://www2.ed.gov/about/offices/list/ocr/docs/tix_dis.html).
[18] Title IX, Civil Rights Division, US Department of Justice. (https://www.justice.gov/crt/title-ix).
[19] US Cabinet. (https://www.whitehouse.gov/administration/cabinet/).
[20] Ostling, A., Sargeant, H., et al., The Cambridge Law Corpus: A Dataset for Legal AI

Research.(https://arxiv.org/abs/2309.12269).
[21] Mahari, R., Stammbach, D., et al., LePaRD: A Large-Scale Dataset of Judges Citing Precedents. (https://arxiv.org/abs/2311.09356).
[22] Li, H., Shao, Y., et al., LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset. (https://arxiv.org/abs/2310.17609).
[23] Pizarro, F., Carenini, G., Neural Multimodal Topic Modeling: A Comprehensive Evaluation. (https://arxiv.org/abs/2403.17308).
[24] Wu, X., Nguyen, T., et al. , FASTopic: A Fast, Adaptive, Stable, and Transferable Topic Modeling Paradigm. (https://arxiv.org/abs/2405.17978).
[25] Gunther, M., Ong, J., et al., Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents. (https://arxiv.org/abs/2310.19923).
[26] Devlin, J., Chang, M, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (https://arxiv.org/abs/1810.04805).
[27] Dodge, J., Sap, M., et al, Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. (https://arxiv.org/abs/2104.08758).
[28] Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (https://arxiv.org/abs/2203.05794).
[29] Sia, S., Dalmia, A., et al., Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! (https://arxiv.org/abs/2004.14914).
[30] Mclnnes, L., Healy, J., et al. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (https://arxiv.org/abs/1802.03426).
[31] Maaten, L., Hinton, G., Visualizing Data using t-SNE. (https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf).
[32] Malzer, C., Baum, M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. (https://arxiv.org/abs/1911.02282).
[33] Ester, M., Kriegel, H., et al., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. (https://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf).
[34] Angelov, D. Top2Vec: Distributed Representations of Topics. (https://arxiv.org/abs/2008.09470).
[35] Le, Q., Mikolov, T. Distributed Representations of Sentences and Documents. (https://arxiv.org/abs/1405.4053).
[36] Jiao, Y., Zhong, M., et al., Instruct and Extract: Instruction Tuning for On-Demand Information Extraction. (https://arxiv.org/abs/2310.16040).
[37] Chang, E. SocraSynth: Multi-LLM Reasoning with Conditional Statistics. (https://arxiv.org/abs/2402.06634).
[38] Yoshimaru, N., Okuma, M., et al. AsyncMLD: Asynchronous Multi-LLM Framework for Dialogue Recommendation System. (https://arxiv.org/abs/2312.13925).
[39] Lu, L., Chen, S., et al., LLM Discussion: Enhancing the Creativity of Large Language Models via Discussion Framework and Role-Play. (https://arxiv.org/abs/2405.06373).
[40] Command-R Documentation, Cohere. (https://docs.cohere.com/docs/command-r#model-details).
[41] Touvron, H., Martin, L., et al., Llama 2: Open Foundation and Fine-Tuned Chat Models. (https://arxiv.org/abs/2307.09288).

[42] Abdin, M., Jacobs, S., et al., Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. (https://arxiv.org/abs/2404.14219).
[43] Mesnard, T., Hardin, C., et al., Gemma: Open Models Based on Gemini Research and Technology. (https://arxiv.org/abs/2403.08295).
[44] Douze, M., Guzhva, A., et al., The Faiss library. (https://arxiv.org/abs/2401.08281).
[45] Johnson, J., Douze, M., et al., Billion-scale similarity search with GPUs (https://arxiv.org/pdf/1702.08734).
[46] Feng, S., Balachandran, V., et al., FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge. (https://arxiv.org/abs/2305.08281).
[47] Jiang, A., Sablayrolles, A., et al., Mistral 7B. (https://arxiv.org/abs/2310.06825).
[48] Pan, J., Wang, J., et al., Survey of Vector Database Management Systems. (https://arxiv.org/abs/2310.14021).
[49] Gao, Y., Xiong, Y., et al., Retrieval-Augmented Generation for Large Language Models: A Survey. (https://arxiv.org/abs/2312.10997).
[50] Han, Y., Liu, C., et al., A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge. (https://arxiv.org/pdf/2310.11703).