Scholarly Question Answering using Large Language Models in the NFDI4DataScience Gateway

Hamed Babaei Giglou*[1][0000−0003−3758−1454], Tilahun Abedissa Taffa⋆[2,3][0000−0002−2476−8335], Rana Abdullah*[3][0009−0000−2652−5129], Aida Usmanova*[2][0009−0000−0124−8727], Ricardo Usbeck[2][0000−0002−0191−7211], Jennifer D'Souza[1][0000−0002−6616−9509], and Sören Auer[1][0000−0002−0698−2864]

[1] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{hamed.babaei,jennifer.dsouza,auer}@tib.eu
[2] Artificial Intelligence and Explainability, Leuphana Universität Lüneburg, Lüneburg, Germany
aida.usmanova@stud.leuphana.de, ricardo.usbeck@leuphana.de
[3] Semantic Systems, Universität Hamburg, Hamburg, Germany
{tilahun.taffa,rana.abdullah}@uni-hamburg.de

Abstract. This paper introduces a scholarly Question Answering (QA) system on top of the NFDI4DataScience Gateway, employing a Retrieval Augmented Generation-based (RAG) approach. The NFDI4DS Gateway, as a foundational framework, offers a unified and intuitive interface for querying various scientific databases using federated search. The RAG-based scholarly QA, powered by a Large Language Model (LLM), facilitates dynamic interaction with search results, enhancing filtering capabilities and fostering a conversational engagement with the Gateway search. The effectiveness of both the Gateway and the scholarly QA system is demonstrated through experimental analysis.

Keywords: Scholarly Question Answering · Federated Search · Retrieval Augmented Generation · Large Language Models · NFDI4DS Gateway

1 Introduction

With recent advances in artificial intelligence (AI), decision-making has gradually shifted from rule-based systems to machine learning and deep learning-based developments [11]. This paradigm shift has changed how we approach information retrieval and Question Answering (QA) systems, including Scholarly QA. Scholarly QA systems answer natural language questions over bibliographic data sources [2,26]. Notably, scholarly resources appear in different bibliographic repositories. To narrow down the answer search space - federated search comes into play. A federated search platform enables one to navigate the vast landscape of scholarly resources available across multiple databases and repositories [8].

⋆ Equal contribution

Furthermore, federated search aggregates information from multiple sources to provide a comprehensive and holistic view of relevant resources. The efficacy of faceted search in scholarly-based filtering has been well-demonstrated [15], paving the way for robust systems employing federated search methods.

Adhering to FAIR principles [29] in managing research data, initiatives like the NFDI4DataScience[4] (NFDI4DS) consortium have emerged as a collaborative endeavor designed to support researchers throughout the entire research data life cycle, ensuring their practices align with the FAIR principles. The NFDI4DS Gateway, as a part of the NFDI4DS vision [22], includes a federated search. The Gateway - a unified and intuitive search interface that enables users to query various scientific databases such as DBLP, Zenodo, and OpenAlex. The overall aim of the NFDI4DS Gateway is to design an entry point that categorizes and summarises multiple search results (such as researchers, publications, machine learning models, and benchmark results) such that practitioners and researchers gain a swift overview of existing contributions [27].

Retrieval Augmented Generation (RAG) harnesses the power of advanced natural language processing (NLP) techniques to improve the quality and relevance of responses to user queries. Integrating Large Language Model (LLM)-based components is at the core of RAG's functionality. LLMs are the backbone of RAG's response generation process, leveraging extensive training on large text to understand and generate human-like responses. RAG-based scholarly QA systems can seamlessly integrate with federated search to improve the process of filtering and selecting scholarly resources in the context of scholarly research [12]. Therefore, on top of the NFDI Gateway, we built a RAG-based scholarly QA system. The RAG retrieval component of the system scans the retrieved resources to identify the top-N most relevant documents based on the user's question. After placing the resources, the scholarly QA uses an LLM (Large Language Model) to extract correct answers to the user's questions directly from the selected documents. By seamlessly integrating the RAG-based scholarly QA with the Gateway, users can efficiently filter through vast amounts of scholarly content, enabling more targeted and productive research efforts.

This approach aims to enhance the user experience, fostering more intuitive and tailored engagement with available information, ultimately contributing to more effective and nuanced research outcomes. Furthermore, a detailed analysis of our experiments is framed as two main research questions (RQs).

- **RQ1:** To what extent does the federated search implemented in the Gateway achieve optimal performance?
- **RQ2:** How does integrating the Scholarly QA on top of the Gateway improve the retrieval of relevant search results?

Our main contributions are twofold:

[4] https://www.nfdi4datascience.de/

- the NFDI4DS Gateway analysis and completeness of federated search evaluation through information retrieval metrics.
- A scholarly QA system based on RAG on top of the Gateway.

The source code can be found at https://github.com/semantic-systems/nfdi-search-engine-chatbot.

2 Related Works

Data management is a multi-step process that involves obtaining, cleaning, and storing data to allow accurate analysis and produce meaningful results. As an example of this, the Open Research Knowledge Graph [24,4] is an infrastructure for the production, curation, publication, and use of FAIR scientific information with the ultimate goal of providing swift knowledge management within the scientific domain by the digitalization of scholarly articles in the form of the knowledge graph. On the other hand, the federated search [30], as they involve the efficient retrieval of information from multiple data sources, play an essential role in data management as it helps in optimizing the use of data and deriving valuable insights from the data. As shown in [3,11] work, the researchers face a flood of papers that hinders the discovery of necessary knowledge, as a result of this, [11] trained models to identify challenges and directions across the corpus by a dedicated search engine.

Federated search [23] serves as a crucial tool for managing data within scholarly articles, enabling the retrieval of information from diverse sources through a search application constructed on top of one or more data sources [8]. A federated search facilitates information retrieval from multiple scholarly sources, demonstrating remarkable efficacy across various fields, particularly in scientific data management. Shokouhi et al. [23] outlined the challenges inherent in federated search within scholarly domains, delineating three significant hurdles: retrieving relevant documents, identifying suitable collections necessitating knowledge representation and unifying results from multiple sources. Similarly, Kumar et al. [10] dived into how federated search helps libraries and other institutions with a valuable tool to explore various fields and articles. Furthermore, Kirstein et al. [9] introduced *Piveau* as a comprehensive open data management solution grounded in semantic web technologies. Leveraging a spectrum of standards prevalent in the semantic web, such as RDFs and DCAT, this standardization via the semantic web overcomes limitations in search capabilities, ensuring superior quality information retrieval.

The Scholarly QA work in [12] proposes a QA model that extracts question-related full-text scientific articles using an LLM-based retrieval agent and generates answers using RAG techniques. [26] has explored Knowledge Graph QA using an LLM in a few-shot setting for handling bibliographic questions. NLQxform [28] introduces a natural language interface for directly querying the DBLP

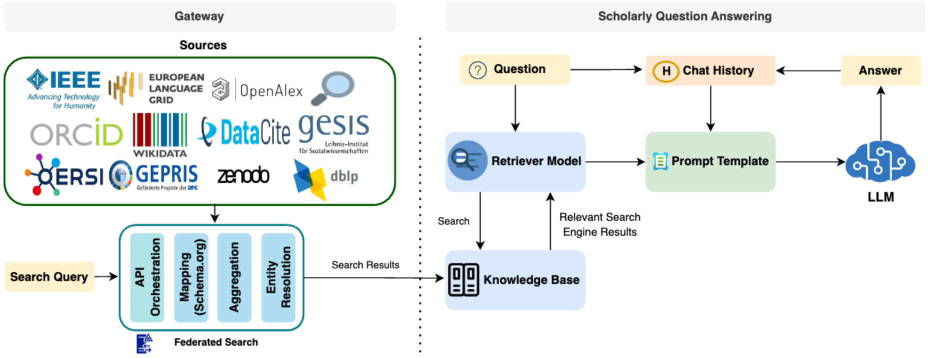

Fig. 1. A functional view of the NFDI4DS Gateway architecture with scholarly QA application.

by automatically translating questions into SPARQL queries. Unlike these Scholarly QAs, we introduce RAG-based Scholarly QA.

3 Methodological Framework

The NFID4DS Gateway performs a federated search through various data store APIs. Subsequently, the search results will be indexed into the QA system to allow users to find acquired information via chat. The architectural representation of the Gateway with a scholarly QA system is illustrated in Figure 1.

3.1 The Gateway – Federated Search

The Gateway conducts federated searches across diverse data stores, generating results that humans can easily interpret. The following key components underpin its functionality:

1. Keyword search across data stores (API Orchestration).
2. Grouping results using a faceted taxonomy (Mapping and Aggregation).
3. Deduplication of results (Entity Resolution).

In the following, we will delve into each of these components in detail, explaining their functions and contributions to the overall functionality of the Gateway.

1) API Orchestration. It uses a one-search-box interface to obtain user keywords, and the search results are expressed in a one-result-list-only manner. It subsequently employs federated search using ad-hoc based searches through 11 open-source scholarly repositories, i.e., DBLP[5], OpenAlex[6], CORDIS[7], Euro-

[5] https://dblp.org/
[6] https://openalex.org/
[7] https://cordis.europa.eu/

pean Language Grid (ELG)[8], GEPRIS[9], GESIS[10], ORCiD[11], RESODATE[12], WIKIDATA[13], IEEE[14], and Zenodo[15]. Among these repositories, DBLP, OpenAlex, IEEE, GESIS, RESODATE, WIKIDATA, and Zenodo provide research resources like publications, datasets, software, etc. GEPRIS provides Deutsche Forschungsgemeinschaft (DFG) funded projects. Likewise, CORDIS is a primary source for projects financed by the European Union (EU) Commission. ELG is a platform that provides multi-lingual, cross-lingual, and mono-lingual language technologies in the EU. Unlike the others, ORCiD delivers a unique, persistent, researcher-owned, and controlled digital identifier that distinguishes researchers uniquely.

2) Mapping and Aggregation. The Gateway interacts with various data source APIs, including SPARQL endpoints; the retrieved results often have different structures. For example, while one source refers to an author of a publication as 'author', another refers to them as 'creator', and terminology differences extend to scholarly resources such as datasets, which are referred to as 'corpus' in one source and 'dataset' in another, such as Zenodo. To resolve the naming variations, we have developed a systematic approach based on customized faceted taxonomy from schema.org[16] that harmonizes and aggregates heterogeneous results from API orchestration. This faceted taxonomy acts as a unifying framework that allows us to map the different terminology and structures found in other data sources, thereby coherently facilitating the aggregation and presentation of search results.

The faceted taxonomy based on schema.org is defined to represent different entities found in data sources. These schema.org classes encompass information including organizations, individuals, authors, creative works (articles, datasets, projects, software applications, learning resources, and media objects), and their respective attributes. In particular, the Author and Person classes encapsulate attributes related to individuals who contribute to creative works, while the Organization class encapsulates attributes specific to organizational entities. In addition, the `CreativeWork` super class serves as a foundation for various entities, providing common attributes such as `abstract`, `author`, and `datePublished` inherited by its subclasses. Each class within schema.org contributes to a structured representation of data entities, facilitating organization, interoperability, and standardized data handling within the Gateway.

[8] https://live.european-language-grid.eu/
[9] https://gepris.dfg.de/
[10] https://www.gesis.org/en/home
[11] https://orcid.org/
[12] https://resodate.org/
[13] https://www.wikidata.org/
[14] https://www.ieee.org/
[15] https://zenodo.org/
[16] https://schema.org/

3) Entity Resolution. Following the initial mapping of the publications, researchers, and other resources using the schema.org taxonomy, it becomes necessary to identify and merge duplicate objects within the results. To accomplish this task, we leverage the DEDUPE model [6], which employs machine learning techniques, specifically fuzzy matching, deduplication, and entity resolution, to handle structured data effectively. Later, the DEDUPE model can be fine-tuned on a custom dataset comprising positive and negative samples, thus enabling the model to differentiate between genuine duplicates and distinct entities.

For publication deduplication, the DEDUPE model is trained on a set of attributes, i.e., Digital Object Identifier (DOI), title, author list, abstract, and publication date for publication identification by clustering objects based on similarity scores calculated across attributes. Subsequently, within each cluster, objects that exceed the predefined similarity threshold are merged to form a unified entity, thus resulting in a set of deduplicated records. Later, the resulting records are sorted based on relevancy score using BM25Plus. BM25Plus is a variant of BM25 (Best Match) [20] ranking algorithm, introducing additional term weighting factors to enhance the ranking.

3.2 Scholarly Question Answering

As shown in Figure 1, our RAG-based [13] scholarly QA has two components: (i) a retriever that returns top-K relevant passage to the user's question and (ii) a generator LLM that generates a human-like response based on a given context from the retriever to a user question.

Retriever. The retrieval model uses a user question as a query to explore relevant information from a knowledge base. The knowledge base comprises a set of documents retrieved per search query through the Gateway. The retriever model operates in three sequential steps:

1. Step 1: The preprocessing knowledge base of search results to obtain a set of documents combined textual data by combining the key-value dictionary per obtained search result.
2. Step 2: The retriever model extracts embeddings for the documents and indexes them within the knowledge base.
3. Step 3: Given a specific question, the retriever model extracts embeddings and computes cosine similarity with the knowledge base, thereby retrieving the top-K appropriate relevant documents to answer the question

We opted for an ensemble retriever model. This ensemble accompanies techniques such as TF-IDF [21], SVM, and KNN retrievers with the Sentence-BERT [19] model serving as the foundational framework. Per the user question, the ensemble retriever queries retriever models to obtain their results; next, it ranks them using each retriever's weights to obtain the final documents most similar to the query. In our retriever collection, the SVM is being trained with the query as a positive class and the rest of the knowledge base documents as negative using sentence-BERT embeddings; next, based on the positive class probability, the documents are ranked and obtain top-k items. By integrating diverse retrieval methodologies, our ensemble model aims to capitalize on the strengths of each component, thereby enhancing overall retrieval performance. Upon experimentation, we manually determined the optimal configuration for our ensemble model. Based on our observations, we assigned weights of 0.3 to TF-IDF, 0.3 to KNN, and 0.4 to SVM retrievers by try-and-error analysis.

Generator. As shown in Figure 1, the generator model uses LLM and retriever documents and a prompt template to query LLM to generate a human-like answer to the user questions based on obtained relevant documents from the retriever model. As observed, LLMs showed a great capability for generating human-like responses. However, they might hallucinate and forget the discussion due to the overwhelming information. We provide explicit instructions beside questions and relevant documents, using a predefined prompt template to avoid this. The prompt template enables the scholarly QA to query LLM effectively and answer the user question accurately. The prompt template is described as follows:

> Provide your answers only on the knowledge provided here. Do not use any outside knowledge.
> If you don't know the answer, say that you don't know. Don't try to make up an answer.
> Given the following context, answer the below question:
>
> {context}
>
> Question: {question}
> Helpful Answer:

In the prompt template, {*context*} is the placeholder for retriever model results, and {*question*} is the user question. To account for follow-up questions, we have used conversation buffer memory that keeps track of chat history, consisting of previous questions and answers within five previous conversations. The follow-up questions can reference past chat history, e.g., "What is the open research knowledge graph?" followed by "How to use it?" Such queries challenge direct retriever similarity-based searches, including ensemble retriever models. We provided the chat history for LLM in the prompt template by adding the history questions and answers to the end of retrieval model outputs at {*context*} placeholder. As an LLM, we use GPT-3.5 [16] with the LangChain framework [5] for implementation.

4 Evaluation

4.1 Evaluation Dataset

This section outlines the procedures for constructing the dataset for both the Gateway and scholarly QA evaluations.

Constructing Queries for Assessing the Gateway Performance. The comparison feature of ORKG empowers researchers to construct comprehensive comparisons [25] among scholarly articles spanning diverse domains. A pivotal aspect of this feature is the inclusion of human-generated comparisons. In the evaluation of federated search, we focused on the comparison titles at ORKG, crafted by the researchers themselves. Consider a scenario where a user aims to formulate a comparison for “ontology learning from text” and utilizes the Gateway to gather relevant papers and sources for their study. When a user queries the title on the Gateway, a user can easily use the documents obtained to construct an ORKG comparison for “ontology learning from the text” as shown in https://orkg.org/comparison/R186047. So, comparison titles can be used as a query to study the Gateway’s performance in finding relevant documents for researchers.

Through this process, we obtained 1,235 unique comparisons from ORKG as of *February 2nd, 2024*, spanning 161 research fields. Among the obtained research fields, we selected 27 research fields related to AI and data science. Consequently, we identified 316 comparison topics within 27 research fields that fall into the AI and data science category for human annotations to curate titles as a query. Ultimately, we curated a collection of 275 comparison titles for performance analysis of the Gateway and executed queries on the Gateway as of *February 16th, 2024*. The remaining 41 comparison titles we found them inappropriate for querying the Gateway.

Generating Scholarly QA Datasets. We designed a systematic approach to generate well-suited questions tailored to search results. The questions are designed to simulate what questions users ask while using the Gateway. We constructed the AI-QA dataset using GPT-4 [17] and the Comparison-QA dataset using ORKG comparisons. For the AI-QA dataset, we employed k-means [7] clustering methodology on retrieved documents per query, enabling us to efficiently organize the data for generating questions. For search result sets containing more than 50 entries, we applied a clustering number of 10, and for result sets with fewer than 50 entries, a clustering number of 5 was considered appropriate. Search results with less than 5 entries were not included in question generation. Subsequently, we employed GPT-4-Turbo [17] to generate two appropriate questions per cluster using a predefined prompt template. The prompt template is defined as follows:

> The task is to generate questions based on the provided information.
> Given a list of texts, generate only two questions, no more than two.
> Make questions variant.
> The questions should imitate what a user might look for in the given documents.
>
> Return questions as a Python list.

Documents:
{documents}

This approach proves advantageous in generating questions for scholarly QA evaluation as it relies on documents already recognized for question generation. However, in the evaluation phase, the retriever model gathers search results similar to those of the questions, which the LLM later uses to generate answers. Following the question generation step, we acquired a total of **3,298** questions across 1,651 clusters for scholarly QA evaluations, where we consider each clustering per question as a ground truth.

Since the ORKG comparison is aimed to allow researchers to compare contributions of different articles based on predefined properties such as "research problem" or "model". For Comparision-QA, we used comparison properties as questions using the following standard template:

In the paper "{paper}", what is the {property}?

We considered 275 comparison titles to query the Gateway to obtain federated search results; for the 275 ORKG comparisons comprising 2,395 papers, only 184 were retrieved by the Gateway. So, we used 184 papers and their properties to construct questions, and values for the property per paper in the comparison were considered as answers. In the end, a total of 1,354 questions were constructed.

The overview of the datasets is presented in Table 1.

Table 1. Statistics for the number of search queries (Query), number of comparison papers (Comparison Papers), number of papers from ORKG comparison that are being covered in search results (ORKG Coverage), and comparison specific questions (Comparison-QA).

Query	AI-QA	Comparison Papers	ORKG Coverage	Comparison-QA
275	3,298	2,303	184	1,354

4.2 Evaluation Metrics

Gateway Evaluation Metrics. In evaluating the performance of the Gateway, we employed multiple approaches, primarily focusing on response time, number of retrieved documents, and relevancy scores. The response time analysis serves as a critical metric in assessing the efficiency and responsiveness of the Gateway. Another key aspect of our evaluation involved analyzing the number of documents retrieved by the Gateway in response to user queries. This metric provides valuable information about the comprehensiveness and effectiveness of the search results generated by the system. To further refine our evaluation, we calculated relevancy scores per retrieved document similarity to the search query based on varying thresholds and representations such as sentence-BERT, TF-IDF, and BM25 [1]. With sentence-BERT sentence embeddings, TF-IDF, and BM25 scores, we calculated cosine similarity between documents and queries for all metrics to get relevancy scores.

Scholarly QA Evaluation Metrics. In AI-QA, we utilized question clusters as answers, while in comparison-QA, property values were employed as answers. Subsequently, we assessed performance using n-gram overlap specific metrics like ROUGE [14] (Recall-Oriented Understudy for Gisting Evaluation) and BLEU [18] (Bilingual Evaluation Understudy), focusing specifically on ROUGE-1, ROUGE-L, and BLEU-1 as our evaluation criteria. Because LLMs generate responses based on their comprehension, they might deviate from the ground truth text, making evaluation with metrics like ROUGE and BLEU difficult. Consequently, incorporating similarity scores into the assessment process can offer further insights into their proficiency in capturing subtle language nuances. We used the BERTScore – a sentence-BERT average cosine similarity metric as an evaluation. Furthermore, as the Comparison-QA dataset poses challenges with answers often appearing within the paper context rather than solely in abstracts and titles, we opted for the Exact Match score as another evaluation metric only for this dataset.

4.3 Results

Gateway and Scholarly QA Results. The performance of the Gateway has been assessed by considering factors such as its response time, the number of documents retrieved, and the relevance of those documents. The Gateway performances are reported in Figure 2 and Figure 3. The results for scholarly QA evaluation, employing various metrics, are reported in Table 2. We identified 432 questions without answers for AI-QA, while we obtained 26 questions without answers for Comparison-QA. This happened due to the input limitation of GPT-3.5. Hence, we excluded these questions from evaluations.

Table 2. Evaluation results of the scholarly QA using AI-QA and Comparison-QA datasets, showcasing ROUGE, BLEU, BERTScore, and Exact Match scores for the RAG-based scholarly QA development.

Dataset	**ROUGE-1**	**ROUGE-L**	**BLEU-1**	**BERTScore**	**Exact Match**
AI-QA	4.21	2.92	38.94	36.81	-
Comparison-QA	6.82	6.10	3.10	26.96	13.93

RQ1: [Gateway] To what extent does the federated search implemented in NFDI4DS achieve optimal performance? We address this question by analyzing the findings presented in Figure 2 and Figure 3. Ultimately, for a search platform, it is essential to retrieve relevant results while maintaining

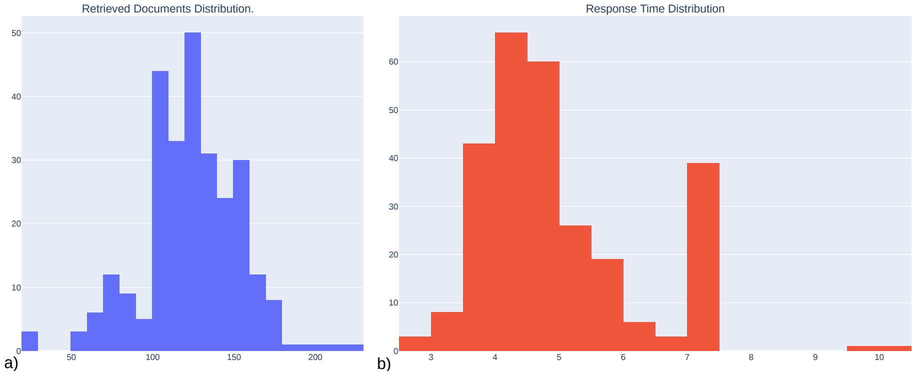

Fig. 2. Gateway retrieved documents distribution is presented in the left figure. The x-axis represents the number of retrieved documents, and the y-axis the number of queries. The right figure represents the response time distribution, with the x-axis as a response time in seconds and the y-axis as the number of queries.

a fast response time across various queries. The analysis of response time and retrieved documents status in Figure 2 for 275 search queries showed that the federated search is capable of obtaining **123** documents on average within an average response time of **4.93** seconds. Notably, slow performance is observed in the search query of the "Kinect human activity recognition dataset" with approximately 10 seconds response time and search results of 169 documents. Similarly, for the "Motion Capture system" search query, we obtained 227 documents within 4.3 seconds. This shows that depending on different search keywords and how complex the query is, it may result in sacrificing response time. In general, according to Figure 2, the distribution analysis indicated that the number of retrieved documents follows a *normal* distribution, while the distribution of response time is *positively skewed*. This highlights the significant performance of the Gateway in terms of response time and document retrieval.

We calculated cosine similarities with three metrics to analyze the retrieved documents' relevancy. We set relevancy thresholds to see how many queries with their corresponding documents are considered very relevant to each other. The relationship between the relevancy threshold and the number of retrieved documents is depicted in Figure 3, indicating a decrease as the threshold increases. The TF-IDF metric generates the highest similarity scores between documents and queries, albeit focusing primarily on token frequency rather than semantic understanding. BM25, an improvement upon TF-IDF, proves particularly effective for information retrieval tasks, displaying a different score distribution with numerous low similarity scores. Despite this, BM25 still identifies certain documents as highly relevant (with similarity above 0.3) for specific queries. Conversely, sentence-BERT initially achieves the highest average recall but drops to zero at a threshold of 0.8. Comparatively, BM25 and sentence-BERT yield similar results, implying that capturing nuanced semantics may not be crucial for

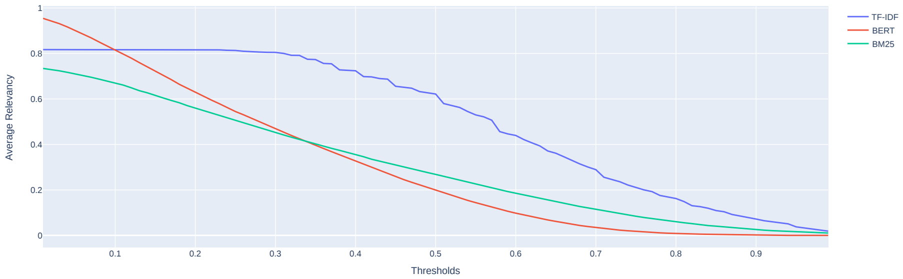

Fig. 3. Gateway retrieved documents relevancy w.r.t search query analysis using TF-IDF, BM25, and sentence-BERT embeddings for similarity measurement and different thresholds in the range of [0.0, 0.99]

retrieving relevant articles; instead, identifying standard terms and phrases appears more pivotal. Evaluating the optimal threshold of 0.3, TF-IDF emerges as the optimal ranking model. The overall relevancy analysis across different thresholds indicates that the Gateway effectively retrieves search results based on keyword search but struggles with semantic retrieval. However, setting the threshold to 0.3 demonstrates approximately 50% semantic similarity among documents, highlighting the Gateway's proficiency in identifying relevant documents from keyword and semantic perspectives.

RQ2: [Scholarly QA] How does integrating the Scholarly QA on top of the Gateway improve the retrieval of relevant search results? We address this question by analyzing the results presented in Table 2 for both automated constructed Comparison-QA and AI-QA datasets. According to the ROUGE-1 metric, unigrams overlap between the developed QA-generated responses and existing answers. This overlap is more significant for Comparison-QA (6.82%) than for AI-QA (4.21%). Similarly, when considering ROUGE-L, which measures the Longest Common Subsequence, the overlap for Comparison-QA (6.10%) surpasses that of AI-QA (2.92%). However, despite the QA's promising BLEU-1 score of 38.94% on the AI-QA dataset, its performance on the Comparison-QA dataset is lacking. This suggests that the developed QA responses align more closely with the clustered documents, which are the ground truth in our AI-QA dataset.

It is essential to note that both the ROUGE and BLEU metrics have limitations when applied to LLM-based generations. This is because LLM-generated responses may exhibit variations that mimic human-like responses, making it challenging for these metrics to evaluate their quality accurately. Still, they show how much of the generated text is similar to ground truth. Nevertheless, we reported a BERTScore of 36.81% for the AI-QA dataset and 26.96% for the

Comparison-QA dataset. These obtained BERTScore results suggest that the quality of the scholarly QA's responses, particularly in terms of semantic similarity to ground truth references, varies significantly between the two datasets. As mentioned earlier, the variation between the two datasets was expected since the Comparison-QA mostly extracted humans from the whole body of the paper rather than only the title and abstract.

We computed the exact match for Comparison-QA, revealing a 13.93% match between the ground truth and the QA-generated text. This highlights the scholarly QA's proficiency in recognizing relevant information, mainly when it appears in the search results. To the best of our knowledge, there is no other baseline system or scholarly QA system available to which we can compare.

5 Limitations and Future Directions

This section discusses the limitations encountered in the implementation of the scholarly QA model and outlines potential future directions for addressing these shortcomings.

Inadequate Availability of Comparison-QA Dataset Answers. The scholarly QA's performance is hindered by the frequent unavailability of answers to the Comparison-QA answers in search results, resulting in suboptimal performance. Addressing this limitation requires an extensive collection of queries from ORKG comparisons. Another limitation arises from the lack of diversity in the questions, as the current methodology employs a single template for forming questions on this dataset.

Suboptimal AI-QA Dataset Generation. The AI-QA dataset, generated from clustered search results, sometimes yields many documents per cluster. Thus, an optimal clustering method is necessary to manage the data effectively. Additionally, soliciting human feedback on the generated questions is crucial for refining and enhancing the dataset's quality. In future works, it is helpful to have a small human-generated dataset to justify the evaluation's validity further.

Exploring Diverse LLMs. Future research should focus on exploring a more comprehensive range of LLMs within scholarly QA to study their diversity and identify more optimal models for scholarly documents. This endeavor necessitates dataset curation tailored explicitly to the Gateway results.

6 Conclusion

In this work, we present an interactive scholarly QA system based on the RAG approach on top of the NFDI4DataScience Gateway search results, facilitating user interaction with a wealth of data. Subsequently, we automatically evaluated both the Gateway and scholarly QA using an automatically constructed dataset. The analysis indicates that as early prototypes, both the Gateway and QA show satisfactory performance. However, there is a need for future work to stabilize both systems and harness data science expertise.

Acknowledgments This work was jointly supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259 and the German BMBF project SCINEXT under grant No. 01lS22070.

References

1. Amati, G.: BM25. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, pp. 257–260. Springer US (2009). https://doi.org/10.1007/978-0-387-39940-9_921
2. Auer, S., Barone, D.A., Bartz, C., Cortes, E.G., Jaradeh, M.Y., Karras, O., Koubarakis, M., Mouromtsev, D., Pliukhin, D., Radyush, D., et al.: The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge. Scientific Reports **13**(1), 7240 (2023), https://www.nature.com/articles/s41598-023-33607-z
3. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.: Towards a knowledge graph for science. In: Akerkar, R., Ivanovic, M., Kim, S., Manolopoulos, Y., Rosati, R., Savic, M., Badica, C., Radovanovic, M. (eds.) Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, WIMS 2018, Novi Sad, Serbia, June 25-27, 2018. pp. 1:1–1:6. ACM (2018). https://doi.org/10.1145/3227609.3227689
4. Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K.E., Vogt, L., Prinz, M., Wiens, V., Jaradeh, M.Y.: Improving access to scientific literature with knowledge graphs. Bibliothek Forschung und Praxis **44**(3), 516–529 (2020). https://doi.org/doi:10.1515/bfp-2020-2042
5. Chase, H.: LangChain (Oct 2022), https://github.com/langchain-ai/langchain
6. Gregg, F., Eder, D.: dedupe (Jan 2022), https://github.com/dedupeio/dedupe
7. Jin, X., Han, J.: *K*-means clustering. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 563–564. Springer (2010). https://doi.org/10.1007/978-0-387-30164-8_425
8. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K.A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Eichner, H., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**(1-2), 1–210 (2021). https://doi.org/10.1561/2200000083
9. Kirstein, F., Stefanidis, K., Dittwald, B., Dutkowski, S., Urbanek, S., Hauswirth, M.: Piveau: A large-scale open data management platform based on semantic web technologies (2020)
10. Kumar, S., Sanaman, G., Ra, N.: Federated search: New option for libraries in the digital era (January 2007)
11. Lahav, D., Saad-Falcon, J., Kuehl, B., Johnson, S., Parasa, S., Shomron, N., Chau, D.H., Yang, D., Horvitz, E., Weld, D.S., Hope, T.: A search engine for

discovery of scientific challenges and directions. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 11982–11990. AAAI Press (2022). https://doi.org/10.1609/aaai.v36i11.21456

12. Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S.G., White, A.D.: Paperqa: Retrieval-augmented generative agent for scientific research. CoRR **abs/2312.07559** (2023). https://doi.org/10.48550/ARXIV.2312.07559
13. Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
14. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
15. Mahdi, M., Ahmad, A., Ismail, R., Kadhim, H., Mohammed, M.: Solution for information overload using faceted search - a review. IEEE Access **8**, 1–1 (06 2020). https://doi.org/10.1109/ACCESS.2020.3005536
16. OpenAI: Chatgpt. https://openai.com/chat-gpt/ (2023), accessed May 5, 2023
17. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023). https://doi.org/10.48550/ARXIV.2303.08774
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135
19. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1410
20. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019, https://doi.org/10.1561/1500000019
21. Sammut, C., Webb, G.I.: TF-IDF. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 986–987. Springer (2010). https://doi.org/10.1007/978-0-387-30164-8_832
22. Schimmler, S., Wentzel, B., Bleier, A., Dietze, S., Karmakar, S., Mutschke, P., Kraft, A., Taffa, T.A., Usbeck, R., Boukhers, Z., Auer, S., Castro, L.J., Ackermann, M.R., Neumuth, T., Schneider, D., Abedjan, Z., Latif, A., Limani, F., Ahmad, R.A., Rehm, G., Khorasani, S.A., Lieber, M.: NFDI4DS infrastructure and services. In: Klein, M., Krupka, D., Winter, C., Wohlgemuth, V. (eds.) 53. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2023, Designing Future - Zukünfte gestalten, Berlin, Germany September 26-29, 2023. LNI, vol. P-337, pp. 919–924. Gesellschaft für Informatik, Bonn (2023). https://doi.org/10.18420/INF2023_103

23. Shokouhi, M., Si, L.: Federated search. Found. Trends Inf. Retr. **5**(1), 1–102 (2011). https://doi.org/10.1561/1500000010
24. Stocker, M., Oelen, A., Jaradeh, M.Y., Haris, M., Arab Oghli, O., Heidari, G., Hussein, H., Lorenz, A.L., Kabenamualu, S., Farfar, K.E., Prinz, M., Karras, O., D'Souza, J., Vogt, L., Auer, S.: Fair scientific information with the open research knowledge graph. FAIR Connect **1**, 19–21 (01 2023). https://doi.org/10.3233/FC-221513
25. Stocker, M., Oelen, A., Jaradeh, M.Y., Haris, M., Oghli, O.A., Heidari, G.s., Hussein, H., Lorenz, A.L., Kabenamualu, S., Farfar, K.E., Prinz, M., Karras, O., D'Souza, J., Vogt, L., Auer, S.: Fair scientific information with the open research knowledge graph. FAIR Connect **1**(1), 19–21 (2023). https://doi.org/10.3233/FC-221513
26. Taffa, T.A., Usbeck, R.: Leveraging llms in scholarly knowledge graph question answering. In: Banerjee, D., Usbeck, R., Mihindukulasooriya, N., Singh, G., Mutharaju, R., Kapanipathi, P. (eds.) Joint Proceedings of Scholarly QALD 2023 and SemREC 2023 co-located with 22nd International Semantic Web Conference ISWC 2023, Athens, Greece, November 6-10, 2023. CEUR Workshop Proceedings, vol. 3592. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3592/paper5.pdf
27. Usbeck, R., Abedissa, T., Veliz, R.A.G., Abdullah, R., Shams, N., Wentzel, B., Chen, Z., Schimmler, S.: NFDI4DS gateway and portal. In: Sure-Vetter, Y., Goble, C.A. (eds.) 1st Conference on Research Data Infrastructure - Connecting Communities, CoRDI 2023, Karlsruhe, Germany, September 12-14, 2023. TIB Open Publishing (2023). https://doi.org/10.52825/cordi.v1i.391
28. Wang, R., Zhang, Z., Rossetto, L., Ruosch, F., Bernstein, A.: Nlqxform: A language model-based question to SPARQL transformer. CoRR **abs/2311.07588** (2023). https://doi.org/10.48550/ARXIV.2311.07588
29. Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The fair guiding principles for scientific data management and stewardship. Sci Data **3**, 160018 (Mar 2016). https://doi.org/10.1038/sdata.2016.18, erratum in: Sci Data. 2019 Mar 19;6(1):6
30. Yang, L., Tan, B., Zheng, V.W., Chen, K., Yang, Q.: Federated recommendation systems. In: Yang, Q., Fan, L., Yu, H. (eds.) Federated Learning - Privacy and Incentive, Lecture Notes in Computer Science, vol. 12500, pp. 225–239. Springer (2020). https://doi.org/10.1007/978-3-030-63076-8_16