Enhancing Health Information Retrieval with Large Language Models: A Study on MedQuAD Dataset

1st Prajwol Lamichhane
School of Computing
University of North Florida
Jacksonville, Florida
0009-0009-5364-7688

2nd Indika Kahanda
School of Computing
University of North Florida
Jacksonville, Florida
0000-0002-4536-6917

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00324

***Abstract*—Given the enormous volume of textual data generated in the healthcare sector, effective and accurate retrieval systems are essential. A major challenge is presented by the explosive growth of scientific publications and medical information. In this study, an advanced pipeline was developed to enhance information retrieval in the healthcare domain. The pipeline has two components: information retrieval and evaluation. The information retrieval component is composed of the retriever, which uses BM25, and the reader, which is powered by a pre-trained Large Language Model. The evaluation component, which uses standardized dataset formats such as SQuAD, provides a framework for evaluating system performance and comparing different parameters. Based on the Cancer category in the MedQuAD dataset, the retriever component showed a strong recall of 0.881 and a Mean Reciprocal Rank score of 0.804, demonstrating its effectiveness in retrieving relevant information and accurate ranking. A Semantic Answer Similarity score of 0.677 for the reader component indicates room for improvement. This work has implications for healthcare providers and the text-mining community working in health information retrieval.**

***Index Terms*—health information retrieval, BM25, pre-trained large language models, MedQuAD**

I. Introduction

In an era of extensive textual information in the healthcare business, the development of efficient and accurate information retrieval systems has become crucial. The healthcare system generates a massive amount of data, which includes electronic health records, scholarly articles, clinical guidelines, and patient-generated information. It is critical for healthcare providers, researchers, and decision-makers to effectively extract important knowledge from this enormous amount of information.

The exponential growth of scientific publications and medical information, as demonstrated by Esther Landhuis' [1] claim that the number of publications increases by 8-9% per year and that approximately 1 million articles are added to the PubMed[1] databases each year, poses a significant challenge in the field of health. An effective and dependable information retrieval system is required to navigate this huge and ever-expanding sea of information. Healthcare practitioners, researchers, and decision-makers would struggle to obtain the most up-to-date and relevant knowledge required for evidence-based decision-making and excellent patient care in the absence of such a system. With Rydning et al. [2] projecting a staggering 36% average annual growth rate for medical information by 2025, the need for a robust information retrieval system becomes even more critical to keep up with the growing volume of valuable research and ensure that healthcare practitioners can access the most relevant and reliable information on time.

In this work, we are developing a Health Information Retrieval (HIR) system, a comprehensive system that incorporates many components to facilitate efficient and accurate retrieval of health-related information, as one promising solution. This pipeline includes several stages, such as collecting data, cleaning, pre-processing, and the deployment of complex algorithms and machine learning models. Deep learning models and cutting-edge techniques are used in this pipeline to increase the functionality and effectiveness of information retrieval in the healthcare domain. It is made up of two key architectures: information retrieval and evaluation.

The information retrieval component consists of a document store, a retriever, and a reader. The retriever component examines accessible texts and selects those that are most relevant to a given query. The selected documents are then processed further by the reader component, which employs advanced deep-learning algorithms to extract top answer choices. Evaluation is a critical component of the health system's information retrieval pathway. It enables us to evaluate the system's quality and performance and compare different techniques. Evaluation data, which is frequently formatted using standardized datasets such as SQuAD [3], serves as a benchmark for measuring our pipeline's effectiveness.

The use of an accurate Information Retrieval pipeline in the health care system has the potential to transform information access, knowledge extraction, and decision-making processes. Medical professionals can make better decisions and deliver better care by efficiently collecting relevant medical publications, clinical guidelines, and research findings. Researchers can gain access to essential resources for evidence-based practice and advancement of the subject.

While our work focuses on the use of this pipeline in the context of cancer information retrieval, it is crucial to note that the pipeline can be modified to serve additional health

[1] https://pubmed.ncbi.nlm.nih.gov/

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00324

domains. This approach has distinct advantages over platforms like Google or ChatGPT [4] as it can be made to utilize the articles exclusively from authentic/reputable resources, such as Cancer.gov [5], ensuring the reliability and accuracy of the indexed information. By building a platform based on trustworthy sources, we aim to create a robust and reliable resource, mitigating the risk of misinformation.

II. Literature Review

Due to the growing need for effective and precise knowledge extraction from enormous amounts of textual material, the fields of information retrieval and question-answering have made enormous improvements in recent years. Particularly, the use of machine learning (ML) models in question-answering and medical information retrieval systems has attracted a lot of interest. Researchers have looked into a range of strategies and methods to improve the functionality of these systems in the healthcare industry and address issues.

A comprehensive review [6] of the application of fuzzy-based machine learning models in medical information retrieval systems for electronic healthcare records was presented by Sengan, Sudhakar, et al. in 2020. The goal is to discuss the challenges and risks of applying ML models in healthcare and to propose alternate approaches. To build predictive models, the research emphasizes collecting the sequence of patient visits and events in electronic health data. The suggested method combines machine and human-derived traits with fuzzy logic to improve model robustness and accuracy. The report also emphasizes how crucial it is to address general security and privacy issues in healthcare, as well as the deployment of ML models. The study outlines possible cognitive computing uses in clinical trials, clinical decision assistance, and EHR conversion. Their main focus was on the need for scientifically proven and trustworthy ML methods in healthcare, while also addressing privacy and security concerns.

In 2022, Sousa, Norberto, Nuno Oliveira, and Isabel Praça et al. [7] offered a solution to the issue of locating relevant information in an enormous number of scientific articles utilizing question-answering approaches. The *Haystack* framework, which enables the application of advanced algorithms to practical use cases, is used to build the proposed system. The system comprises a user-facing web interface and a back-end RESTful API that can fetch scientific publications, review the database summary, and identify potential solutions. The system uses a two-step process, with a reader component that uses the RoBERTa model to extract possible answers and a retriever component that uses TF-IDF to find relevant documents. The system is examined in the disciplines of energy and cybersecurity, showcasing its efficiency in locating accurate and relevant answers to research questions in those sectors. Future work will involve increasing the system's capabilities as well as its effectiveness and scalability.

Izacard and Grave proposed an approach [8] in 2021 for training a question-answering information retrieval system without needing annotated query-document pairs. A reader and a retriever are the two modules that make up the system. The retriever selects relevant passages from a large knowledge source, while the reader processes these passages along with the question to generate an answer. Their work focuses on retriever module training without strong supervision. They propose training the retriever by learning to approximate the attention scores of the reader, assuming that these scores are a good indicator of passage relevance. The reader module's Fusion-in-Decoder model is discussed, along with the design of the retriever and how it is trained. According to experimental results, their technique outperforms current state-of-the-art models in benchmark tests for question answering.

In their research work [9], Asma Ben Abacha and Dina Demner-Fushman offer a brand-new question-answering (QA) strategy based on RQE (Recognizing Question Entailment) for the medical field. The research emphasizes the difficulties of question interpretation and answer extraction tasks and illustrates the difficulties in large-scale information retrieval for answering natural language questions. A QA system and resources designed for actual medical questions are introduced by the authors, together with a collection of 47,457 question-answer pairs from credible medical sources. They evaluate the effectiveness of combining IR models with RQE for question answering and compare logistic regression and deep learning methods for RQE using various data sets. The evaluation's findings suggest that their method outperforms the best official score in the medical task, demonstrating the value of the question's impact on QA and the potential advantages of combining IR and RQE. The paper also discusses trustworthy response sources and potential options for developing the QA technique, such as expanding the question-answer collection and emphasizing focus recognition.

S. Tian et al. [10] highlights the widespread exploration of ChatGPT and LLMs in biomedical and health-related tasks. They report the absence of actual LLM deployment in biomedical and clinical practice due to challenges such as limited transparency of training data, ethical considerations, patient privacy, and biases in AI models [10]. Simialry Q. Jin et al. [11] discuss issues of "hallucination" in responses. They suggest combining LLMs with search engines to enhance information-seeking, despite current limitations.

III. Methodologies

Various steps were followed while working on the development of an IR system in health. We went through the data collection, data cleaning, data pre-processing, pipeline development, and pipeline evaluation for the IR system. We used Haystack[2], which is an open-source LLM-based Python framework for building custom NLP-driven applications, to develop the core components of our information retrieval pipeline shown in Figure 1. The pipeline consists of two components, (1) information retrieval, and (2) evaluation.

A. Information Retrieval

Figure 1 depicts each of the components in the IR system.

[2]https://haystack.deepset.ai/

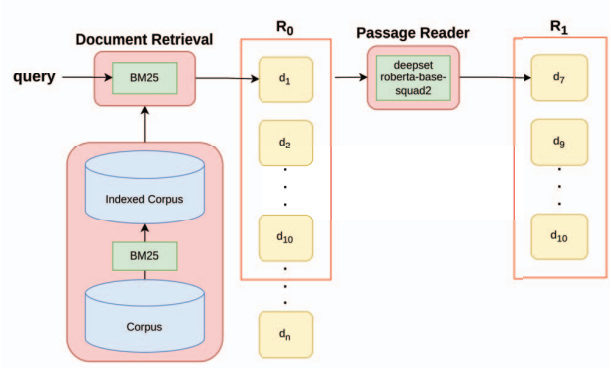

Fig. 1. Multistage Information Retrieval Pipeline

1) Indexing Documents in a Document Store

To start with, a *DocumentStore* was created to store the Documents needed for a question-answering system. While many systems can be used as a documentstore such as Milvus, OpenSearch, Pinecone, *ElasticsearchDocumentStore* was selected as the preferred document store due to its capabilities for sparse retrieval, which offers numerous tuning options, along with basic support for dense retrieval. In order to index the files, the indexing pipeline was used which includes a node called *PreProcessor*. Documents must be preprocessed and divided into smaller portions in Haystack before being indexed in the document store. This is done by the *PreProcessor*. It enables fine-tuning the preprocessing settings to meet particular needs. A variety of initialization settings for the *PreProcessor* include *split_by*, which establishes the unit of splitting, such as split by words, and *split_length*, which establishes the maximum amount of words in each passage. The texts are preprocessed and divided into smaller portions using the *PreProcessor* in the *add_eval_data* method, allowing for more accurate information extraction from the indexed documents. This method can improve the efficiency of searching for and retrieving relevant data during information retrieval or questioning and answering tasks. A default hyperparameter setting provided by Haystack was used on this pipeline.

2) Retriever

To retrieve relevant documents for a given question, a Retriever is employed which evaluates all the available Documents and selects only those that are relevant to the question. While there are various other retrievers like DensePassageRetriever, TableTextRetriever, TfidfRetriever, BM25Retriever [12] is used in the retriever pipeline because it takes into account both the frequency of terms in documents and the length of documents, making it an effective method for syntactic similarity measurement in information retrieval tasks. It is widely regarded as the most suitable Retriever for question-answering systems that rely on IR [12].

3) Reader

After the Retriever selects the relevant documents, a Reader is used to scan the texts in the documents and extract the top answer candidates. Readers use advanced deep-learning models, but they are slower than Retrievers when processing the same amount of text. In this research, a *FARMReader* is used with a base-sized *BERT* [13] and *Roberta* [14] question-answering model named *deepset/bert-base-cased-squad2* and *deepset/Roberta-base-squad2*. Haystack provides BERT-based architectures like BERT, RoBERTa, ALBERT, MiniLM, XLM, etc. The choice of using the FARMReader in conjunction with the Roberta question-answering models, specifically *deepset/bert-base-cased-squad2* and *deepset/Roberta-base-squad2*, is driven by the need to leverage state-of-the-art deep learning techniques for semantic similarity assessment. While the Retriever evaluates the syntactic similarity, the Reader focuses on the semantic similarity of the text.

4) Querying

By combining all these components, the elasticsearch document store, retriever, and reader, a pipeline is created where a query can be given as an input question and the pipeline would result in answers to the query along with the context and the documents where the answer was obtained from. As the retriever and the reader are separate parts, the query process occurs in different steps. When a question is given as input to the pipeline, the retriever results in the top 10 documents that are likely to answer the question. And, the reader's work is to go through the 10 candidate documents obtained by the retriever and find the most relevant answers. The retriever results in the top 5 answers from those documents that are output by the retriever. The 10 and the 5 are hyperparameters that can be tuned according to requirements. Although the final result may have 5 answers, the top answers will have the maximum score to depict that they match the question and the context at the highest level.

B. Evaluation Pipeline

Information retrieval system evaluation is essential for determining how well the system is working and for determining how to make improvements. It enables us to assess the prediction quality of the system, which is crucial in figuring out how effective it is. Additionally, evaluation offers a framework for comparing various systems and measuring their performance against established standards. Evaluation data for question-and-answer systems are frequently formatted using the SQuAD (Stanford Question Answering Dataset) [3] standard. SQuAD is a standard for evaluating the comprehension and question-answering abilities of extractive question-answering systems. The evaluation metrics that we utilized are demonstrated in Figure 2.

Using the *eval()* method on the Haystack pipeline, the evaluation can be performed directly on an existing pipeline with any gold-standard data source. This streamlined method ensures consistency throughout the pipeline implementation stages while providing an evaluation that is quicker and more practical. The actual "gold" labels for the test data are fetched from the document store while the pipeline is being evaluated. Gold labels are the ones that correspond to the dataset's real responses as stored in the elasticsearch *documentstore*. An *EvaluationResult* object contains the evaluation results,

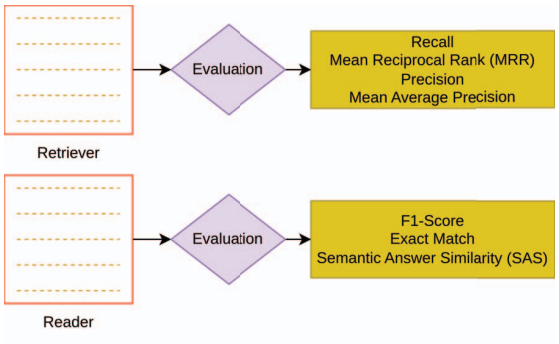

Fig. 2. Information Retrieval Evaluation Metrics

including metrics and reader/retriever node performance. The evaluation results can be accessed and further analyzed using the *EvaluationResult* object. Metrics from the outcomes include Semantic Answer Similarity (SAS), F1 score, and Exact Match. This information can be used to generate an evaluation report for a thorough performance analysis of the pipeline or filtered depending on particular queries of interest.

C. Gold-standard Dataset for Evaluation

When it comes to health, finding a quality dataset is always critical as health is a sensitive field. The correctness and credibility of the dataset were taken into account before working on this research. We used the MedQuAD [9] dataset in this study, which is made up of 47,457 question-answer pairs that were taken from credible NIH websites like cancer.gov, niddk.nih.gov, GARD, and MedlinePlus Health Topics. This dataset includes subjects relating to treatment, diagnosis, side effects, and numerous medical entities like diseases, medications, and tests, as well as 37 different question kinds. For this study, We chose to work with the *cancer* category for our initial effort in a suitable health IR system where we indexed 114 pairs of questions, answers, and context passages on the *documentstore* and evaluated them. This dataset, part of which is shown in Figure 3. contains useful information about the question's type, focus, source, and context.

1) Data Cleaning and Pre-processing

Although a proper set of question-answers was readily available, we required a significant amount of data cleaning and preparation. Since the article sources for each question-answer were mentioned in the MedQuAD dataset, to get updated information, we performed web scraping, and the most recent context from the source links was extracted. Then, filtering out jargon like header/footer information from scraped contents, removing duplicates, manually scraping contents from missing or failed automatic scrapings, etc was carried out.

Similarly, in the data pre-processing, various tasks needed to be performed. When we index the dataset in the IR pipeline, the dataset is expected to be ingested in a SQuAD [3] format. Typically, the SQuAD dataset is offered in JSON format. A single context paragraph, along with any related questions and answer annotations, is represented by each JSON entry.

```
<?xml version="1.0" encoding="UTF-8"?>
<Document id="0000002_1" source="CancerGov"
url="https://www.cancer.gov/types/leukemia/patient/adult-all-treatment-pdq">
  <Focus>Adult Acute Lymphoblastic Leukemia</Focus>
  <FocusAnnotations>
    <UMLS>
      <CUIs>
        <CUI>C0751606</CUI>
      </CUIs>
      <SemanticTypes>
        <SemanticType>T191</SemanticType>
      </SemanticTypes>
      <SemanticGroup>Disorders</SemanticGroup>
    </UMLS>
  </FocusAnnotations>
  <QAPairs>
    <QAPair pid="1">
      <Question qid="0000002_1-1" qtype="treatment">What are the
treatment options for adult ALL?</Question>
      <Answer>Adult ALL can be treated with various approaches
including chemotherapy, radiation therapy, stem cell transplant, and targeted
therapy. The specific treatment plan depends on factors such as the patient's
age, overall health, and genetic characteristics of the leukemia cells.
</Answer>
    </QAPair>
  </QAPairs>
</Document>
```

Fig. 3. A part of MedQuAD Dataset in XML format

The format has fields for the paragraph text ("context"), the question text ("question"), and the answer span(s) with their corresponding character offsets within the context paragraph ("answers"). Although we had access to context, question, and answer in the MedQuAD data, the answer span or the answer start attribute was to be determined. This was challenging because the answer in the MedQuAD dataset was semantically similar to the question and the context, however, our pipeline required syntactic similarity to a specific part of the context as a part of the evaluation. Also, to determine the answer start attribute of the SQuAD format, an exact sentence from the context should be in the dataset as the answer to questions. So, to solve this issue, the following approaches were utilized.

a) Syntactic Similarity Analysis: In this approach, each sentence of the context and the answer was first tokenized, and stop words were filtered using NLTK [15]. Then, a Tfidf vectorizer [16] was used to create a vector out of those tokens. A cosine similarity was obtained between each sentence and the answer from the MedQuAD dataset. The sentence with the highest score was taken as the probable matching sentence and the sentence's starting character index was kept as a value for the answer start attribute of the SQuAD format. However, when the answer sentence obtained using cosine similarity was analyzed, the answer sentence did not look finely convincing. So, the next approach was explored to find convincing answers.

(b) Semantic Similarity Analysis: After syntactic similarity showed mediocre performance, we explored using the BERT-base transformer model [13] for question-answering as a solution. In this case, the context and the questions were utilized from the MedQuAD dataset to find the appropriate sentence that answer the question. However, upon manual inspection, although this approach was better than the syntactic

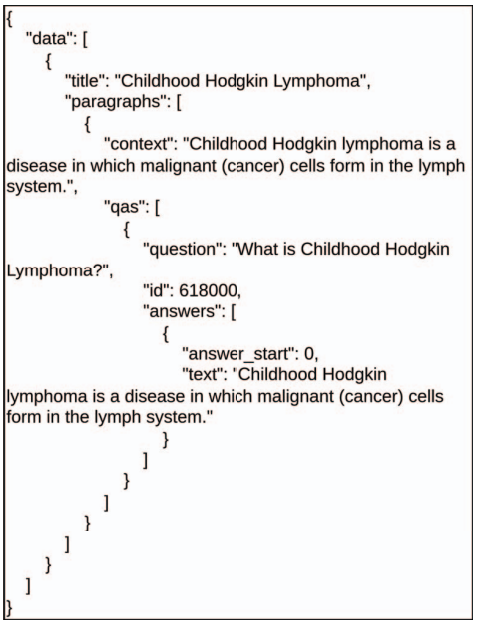

Fig. 4. SQuAD Dataset Format

similarity analysis, it was still imperfect. Thus, manual re-evaluation was necessary to find the answer start attribute.

IV. Results and Discussion

Various indicators were used to assess the performance of the Retriever and the Reader component in the information retrieval component (shown in Table I). In the case of the retriever, the recall, which evaluates the capacity to recover relevant documents, was calculated for both single and multiple retrievals of relevant documents. For all cases, the Retriever scored a recall of 0.881, showing a high degree of success in recovering relevant information. The Mean Reciprocal Rank (MRR) is a measure of the first relevant document's ranking accuracy. The Retriever received an MRR score of 0.804, indicating that the first relevant document was rated quite high in the retrieval results on average. Precision, which measures the accuracy of the retrieved documents, was calculated to be 0.667. This means that around 66.7% of the documents returned were related to the query. The Mean Average Precision (MAP) calculates the average precision of several searches. The Retriever had a MAP score of 0.780, suggesting that it performed well in terms of precision across all of the queries.

Several metrics were used to evaluate the performance of the Reader component. As a baseline, we used the BERT model as an alternative Reader component. The BERT model obtained a SAS score of 0.688, indicating a good level of semantic similarity between the extracted and ground truth responses. This shows that BERT can effectively interpret the context and meaning of the answers. The F1-Score achieved, on the other hand, was 0.260, indicating an acceptable performance in terms of obtaining accurate responses from the recovered documents. Furthermore, the Exact Match score was 0.041,

TABLE I
Performance Metrics of the Retriever and Reader

Retriever Results	**BM25**	
Recall (single relevant document)	0.881	
Recall (multiple relevant documents)	0.881	
Mean Reciprocal Rank	0.804	
Precision	0.667	
Mean Average Precision	0.780	
Reader Results	**BERT**	**RoBERTa**
SAS (Semantic Answer Similarity)	0.688	0.677
F1-Score	0.260	0.449
Exact Match	0.041	0.000

indicating that just a small percentage of the obtained replies matched the ground truth answers exactly. It's worth noting that exact matching may not be required if the question-answer pipeline includes descriptive answers rather than one-word responses. Nonetheless, these results provide a solid baseline.

As shown in Figure 1, our pipeline includes the RoBERTa model, a more powerful large language model. The SAS score for the RoBERTa model was 0.677, indicating a reasonable level of semantic similarity between the retrieved responses and the ground truth answers. This implies that RoBERTa performs well at capturing the context and significance of the responses. The F1-score of 0.449 shows the Reader's overall performance in terms of obtaining accurate answers from the retrieved texts. While this score suggests a reasonable performance, the Exact Match score of 0.0 indicates that none of the extracted responses exactly matched the ground truth answers. However, as previously stated, exact matching may not be required for descriptive question-answer pipelines. This finding, however, highlights the potential for additional adjustments to improve the precision and accuracy of the answer extraction process utilizing RoBERTa.

As noted above, the overall evaluation of the Retriever and Reader components shows that the information retrieval pipeline is effective in retrieving relevant documents and extracting answers from them. However, there is room for improvement, notably in terms of Reader precise match performance. Following are some "challenging" examples for which our pipeline's output was not up to the standard.

Example 1:

Query: What is/are Anal Cancer?
Gold-standard Answer: Anal cancer is a type of cancer that develops in the tissues of the anus.
Proposed System Answer: Women who are HIV-positive also have an increased risk of anal cancer compared with women who are HIV-negative. Studies show that intravenous drug use or cigarette smoking may further increase the risk of anal cancer in patients who are HIV-positive.

The question "What is/are Anal Cancer?" aims to clarify the nature of anal cancer. According to the gold answer, anal cancer is a form of cancer that arises in the tissues of the anus. The pipeline solution, on the other hand, deviates from the proper answer and explains the risk factors for anal cancer in

HIV-positive women. This indicates a clear mismatch between the expected and pipelined answers.

Example 2:

Query: What is/are Primary Myelofibrosis?
Gold-standar Answer: Primary myelofibrosis is a rare bone marrow disorder characterized by the excessive production of fibrous tissue, leading to the replacement of healthy bone marrow with scar tissue.
Proposed System Answer: In primary myelofibrosis, also called chronic idiopathic myelofibrosis, large numbers of blood stem cells become blood cells that do not mature properly (blasts).

"What is/are Primary Myelofibrosis?" queries information on primary myelofibrosis, a bone marrow illness characterized by excessive fibrous tissue creation, resulting in scar tissue replacing healthy bone marrow. This condition is appropriately described by the gold-standard answer. However, our systems' response fails to deliver the correct information, instead mentioning incorrect blood cell maturation in primary myelofibrosis. This disparity exposes the systems' limitations in accurately retrieving the needed information.

V. Conclusion and Future Work

It is crucial to create an effective and precise information retrieval system for the healthcare industry. In this study, we developed an advanced pipeline comprised of information retrieval and information evaluation components using the Haystack framework to handle the difficulty of effectively obtaining healthcare information from a large volume of textual data. The MedQuAD dataset was used to validate the effectiveness of information retrieval and evaluation systems. Both BERT and RoBERTa are effective as the Reader component, with RoBERTa outperforming BERT in terms of accuracy and precision. The outcomes of both models provide useful insights that can be used to refine and improve the Reader component for better response extraction and information retrieval. The use of a standardized dataset for evaluation enabled technique comparison and allowed for a thorough evaluation of system performance. The study ensured robustness and reliability in its findings by leveraging the MedQuAD dataset, enhancing the overall credibility of the research outcomes. Our findings showed promise, proving the system's capacity to retrieve relevant information and extract answers. While there is room for improvement, the study emphasizes the importance of maintaining precision and accuracy to fulfill healthcare practitioners' and researchers' increasing information needs. Several cases of inadequate answers by our system highlight the pipeline's difficulties in reliably predicting or extracting the correct responses. They emphasize the need to improve the pipeline's performance and adopt more precise information retrieval techniques.

The research will be improved further by implementing an optimized data scraping and cleaning approach to ensure that the *documentstore* only contains clean and jargon-free datasets. Also, the current model is pre-trained on general text, which might not be optimized for medical terminology and context. By fine-tuning the model with medical text data, it can better understand medical concepts and terminology, leading to more accurate answers to medical queries. Also, the current research focuses on cancer, the next phase intends to broaden its scope to include a variety of other health concerns, obtaining data from credible sources. Addressing the problems given by the ever-increasing volume of scientific papers and medical information will be a priority, demanding the retrieval system's continuous adaptation and enhancement.

Finally, in our evaluation of the Retriever and Reader components, it's important to note that we did not incorporate specific benchmarks as there are currently no existing benchmarks that accurately capture the intricacies of our information retrieval task. Nonetheless, future work could explore incorporating/adapting existing benchmarks to establish further points of reference and enable more direct comparisons with other information retrieval solutions in the field.

VI. Acknowledgments

We would like to express our sincere gratitude to Nazmul Kazi for his assistance with the data preprocessing.

References

[1] E. Landhuis, "Scientific literature: Information overload," *Nature*, vol. 535, no. 7612, pp. 457–458, 2016.
[2] D. R.-J. G.-J. Rydning, J. Reinsel, and J. Gantz, "The digitization of the world from edge to core," *Framingham: International Data Corporation*, vol. 16, 2018.
[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.
[4] "Chatgpt," https://chat.openai.com/, accessed: 2023-01-15.
[5] "Cancer gov," https://www.cancer.gov/, accessed: 2023-02-06.
[6] S. Sengan, G. Kamalam, J. Vellingiri, J. Gopal, P. Velayutham, V. Subramaniyaswamy *et al.*, "Medical information retrieval systems for e-health care records using fuzzy based machine learning model," *Microprocessors and Microsystems*, p. 103344, 2020.
[7] N. Sousa, N. Oliveira, and I. Praça, "Machine reading at scale: A search engine for scientific and academic research," *Systems*, vol. 10, no. 2, p. 43, 2022.
[8] G. Izacard and E. Grave, "Distilling knowledge from reader to retriever for question answering," *arXiv preprint arXiv:2012.04584*, 2020.
[9] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019.
[10] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau *et al.*, "Opportunities and challenges for chatgpt and large language models in biomedicine and health," *arXiv preprint arXiv:2306.10070*, 2023.
[11] Q. Jin, R. Leaman, and Z. Lu, "Retrieve, summarize, and verify: How will chatgpt impact information seeking from the medical literature?" *Journal of the American Society of Nephrology*, pp. 10–1681, 2023.
[12] S. Wang, S. Zhuang, and G. Zuccon, "Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval," in *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*, 2021, pp. 317–324.
[13] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.
[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[15] "Nltk," https://www.nltk.org/, accessed: 2023-01-31.
[16] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.