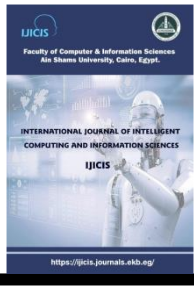
IJICIS, Vol.23, No.1, 84-97
DOI: 10.21608/ ijicis.2023.185241.1245

**International Journal of Intelligent Computing and Information Sciences**

https://ijicis.journals.ekb.eg/



# LONG-FORM OPEN-DOMAIN QUESTION-ANSWERING SYSTEM ARCHITECTURE

Moataz Mohammed*
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
moataz.mohammed@cis.asu.edu.eg

Salsabil A. El-Regaily
Basic Sciences Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
salsabil_amin@cis.asu.edu.eg

Mostafa M. Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mostafa.aref@cis.asu.edu.eg

Received 2023-01-03; Revised 2023-03-03; Accepted 2023-03-07

***Abstract:*** *Question Answering is one of the challenging points of research in natural language processing recently. The problem of automating the answering process for the user's queries became required. So, there were several papers suggested different system architectures for building a question answering systems. In this research paper, we suggest our own system architecture taking into consideration that the input of the system architecture is only the asked question. The suggested system architecture is a long-form open domain question answering that contains mainly two layers. The natural language processing layer which holds the data module and the computing module. This layer is responsible for many operations like pre-processing, preparing, storing the data along with taking the user's question then providing the suitable answer. The dataset of the proposed system has to be documents annotated with questions and answers extracted from these documents. Also, it has to be in SQUAD format. The computing module is a retriever-reader based deep learning model. This model achieves scores: 67% Recall@100 using dense passage retriever model and 67.7% F1 score for reader model. the Interface layer is the second layer which includes the APIs module and the user-interface module. Finally, we will discuss a real time case study for the system.*

***Keywords:*** *Natural Language Processing, Long Form Open Domain Question Answering, Chatbot, Open Domain Question Answering, Semantic Search*

## 1. Introduction

One of the most interesting fields in artificial intelligence these days is the natural language processing (NLP). NLP is a computer science field which is responsible for simulation of the natural language of humans. This imitation process can be performed within collection of tasks like sentiment analysis, Named Entity Recognition NER, word Synonyms, Text Summarization, Text Translation and Question

***Corresponding Author**: Moataz Mohammed

Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: moataz.mohammed@cis.asu.edu.eg

Answering QA. One of the most frequent professions in NLP is the QA task. The QA can be classified in NLP from various angles. The first aspect is the domain-type. The domain in the QA task [1] is either closed-domain question answering CDQA or open-domain question answering ODQA. There are various controversies about the term of "Open" in ODQA which is thought that it refers to questions in any area. This may be a reason. However, the most trusted belief is that it means that the model just needs only one input (question) not two inputs (question with context) to answer any question. While the closed domain means that the input has to include the question with context. Here we are working on the ODQA [2]. Internally, the ODQA may be Open Book or Closed Book. Open Book means that answering the questions is based on extracting the text from data-store which is Informational Extraction (IE) task in NLP. Closed Book is similar to the case of the student when studying for exams. To answer the exam, he needs to remember the information that he studied. So, in the closed book the answer is generated based on the weights of the trained model which called text-generation in NLP. We work in this paper on the information extraction branch (closed book). Second perspective is the data source which is either Knowledge-Based KB or Data-Based DB. KB means the data is stored in structured manner with respect to the meaning about the relationship existence between the entities and the synonyms of entities like Wordnet [3]. DB depends on the source of data and the Deep Learning DL models to learn the relations between the data about the huge number of parameters that store. The work here is based on DB based. The dataset that was collected was about an English Islamic articles' dataset EIAD [4]. Last perspective is during the answer type: factoid answers (Short-Answers) or non-factoid answers (Long-Answers). In our work, we work mainly on the non-factoid answers. The most recent used architectures in developing the QA systems are the Retriever-Reader architectures [5] or Retriever-Generator architectures [6]. This paper is based on the Retriever-Reader architecture. In the following sections, we will discuss the general architecture of building the ODQA system architecture. It starts from collecting the data until using pretrained models to retrieve the relevant documents and answering the asked question on output platform. Finally, we will show a use case for the suggested system architecture.

## 2. Related Work

In recent years, the question answering task has much advancements through the world. There is more than one research paper published during these recent years about the question answering systems. In this research paper, we will discuss some of them especially these ones are retriever-reader or retriever-generator based systems.

### 2.1. End-to-End Training of Multi-Document Reader and Retriever

For retrieval-augmented open-domain question-answering systems that use data from several retrieved documents to generate replies, [7] provides an end-to-end differentiable training strategy. It represents selections for retrieval as latent variables across sets of pertinent texts. Because marginalizing over collections of recovered documents is computationally challenging, the expectation-maximization has been used to approximate it. We update the retriever and reader parameters after iteratively estimating the value of our latent variable (the list of pertinent documents for a particular topic). In contrast to stage-wise training, we predict that end-to-end training permits training signals to flow to the reader and then to the retriever more effectively. This results in a retriever that is able to select more relevant documents for a question and a reader that is trained on more accurate documents to generate an answer. Experiments on three benchmarks datasets demonstrate that our proposed method outperforms all existing approaches of comparable size by 2-3 absolute exact match points, achieving new state-of-the-art results. Our results also demonstrate the feasibility of learning to retrieve to improve answer generation without explicit supervision of retrieval decisions. As a result, a reader who has been taught on more correct documents can produce an answer and a retriever can choose materials that are more pertinent to a topic. Experiments on three benchmark datasets show that our suggested solution achieves new state-of-the-art results, outperforming all other approaches of comparable size by 2-3 absolute exact match points. Our findings also show that it is possible to learn to retrieve in order to enhance answer production without direct monitoring of retrieval choices.

### 2.2. End-to-End Training of Neural Retrievers for Open-Domain Question Answering

Both supervised and unsupervised techniques have been used in recent work on training neural retrievers for open-domain question answering (OpenQA). The best way to combine supervised and unsupervised techniques for neural retrievers is still a matter of debate. We systematically investigate retriever pre-training in this work. First, [8] suggests a method of unsupervised pre-training utilizing the Inverse Cloze Task and masked salient spans, then supervised finetuning with question-context pairings. On the Natural Questions and TriviaQA datasets, this method produces absolute increases of 2+ points above the prior best result in the top-20 retrieval accuracy. We then investigate two methods for training the reader and retriever components in OpenQA models from beginning to end, which vary in how the reader consumes the recovered documents. As a result of our experiments, we are able to produce results that are cutting edge, proving the efficacy of these strategies. We achieve a top-20 retrieval accuracy of 84% on the Natural Questions dataset, which is an improvement of 5 points over the most previous DPR model. Our findings on response extraction are likewise impressive, exceeding more current models like REALM and RAG by 3+ points.

### 2.3. Hurdles to Progress in Long-form Question Answering (LFQA)

Long-form question answering (LFQA) is the process of finding materials that are pertinent to a question and using them to create an answer that is at least one paragraph long. The task formulation raises fundamental challenges regarding evaluation and dataset creation that currently prevent meaningful modelling progress, despite the fact that many models have recently been proposed for LFQA. We demonstrate this in this paper. [9] first create a novel system that uses contrastive retriever learning and sparse attention to attain cutting-edge performance on the ELI5 LFQA dataset in order to illustrate these difficulties. While the public scoreboard places our system at the top, a further investigation reveals some alarming trends: (1) As at least 81% of ELI5 validation questions appear in paraphrased form in the training set, at least ELI5 contains significant train-validation overlap; (2) ROUGE-L is not an informative metric of generated answer quality and can be easily manipulated; and (3) human evaluations used for other text generation tasks are unreliable for LFQA. We provide solutions to each of these problems in the hopes that they would encourage future LFQA research to be more rigorous and produce significant advancements.

## 3. System Architecture

Figure 1 shows block diagram of suggested open domain question answering ODQA system architecture as in [2]. This architecture is Retriever-Reader based QA system architecture. This system architecture has two main layers: NLP Layer which has two main modules like data module and computing module and the Interface Layer which has other two modules such as the Application Programming Interface APIs module and the User Interface module. Each of these modules have their own components. First the data module that has five components like data collection, data preprocessing, data preparing, data vectorization and data storing. The second one is the computing module which has the retriever component and the reader component. The Interface Layer modules are out of scope of the NLP so, we only mention their names and functionality. The APIs module is a container which has any an API related to moving the data into the user interface. The User Interface module refers to the platform which will show the inputs and the outputs of the system on it. In the next sections will explain each of these components in more details.

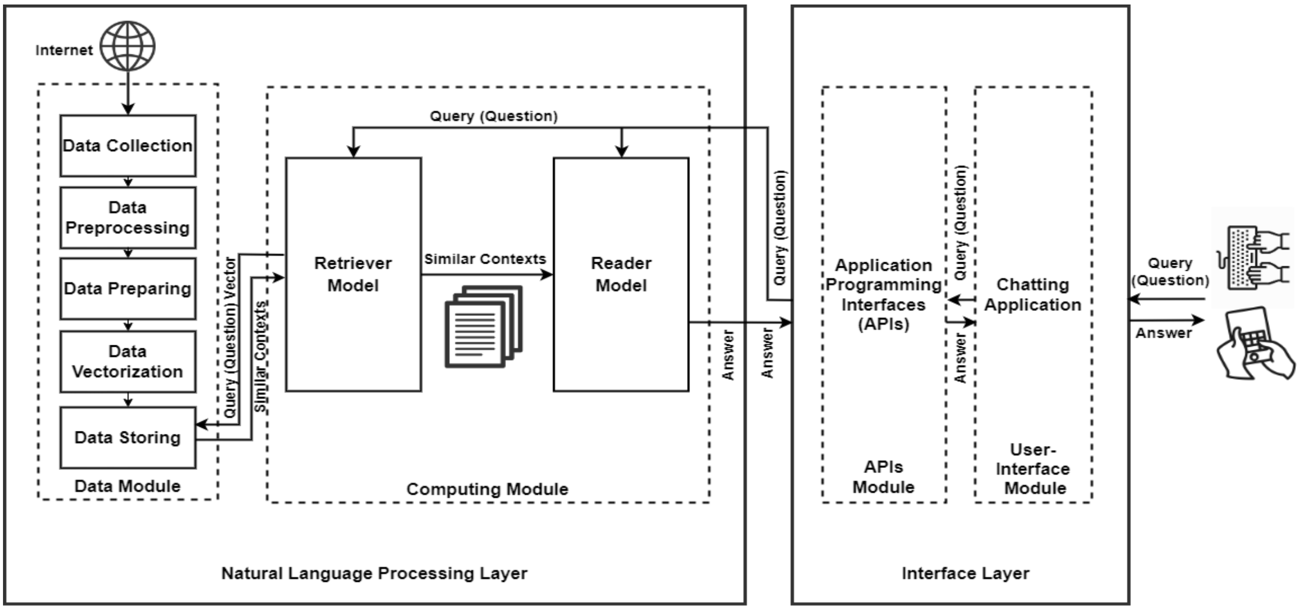



Figure. 1: General Open Domain Question Answering System Architecture

## 3.1. Data Module

The data module is the first module in the NLP layer. This section will discuss each component related to the data module starting from collecting the data until storing it. This module has main five components like data collection, data preprocessing, data preparing, data vectorization and data storing.

### *3.1.1. Data collection*

The technique used for collecting the data in this phase is data scraping. Scraping is a software technique is followed to collect data from the internet. Our own dataset is about an English Islamic Articles Dataset EIAD. EIAD dataset was scraped from most popular genuine Islamic websites like NewMuslims.com [10], IslamReligion.com [11], and IslamQA.com [12]. The process of scraping was about writing our own python script using python packages like Beautifulsoup and Selenium. We had to develop our scrapper by ourselves because of the difference in structure of the former websites. We collected about 1550 articles from IslamReligion.com, 8292 articles from IslamQA.com and 275 articles from NewMuslims.com. EIAD dataset is an initial version which contains about 10000 English Islamic articles categorized into 15 different categories which cover topics like Evidence Islam is Truth, The Benefits of Islam, Beliefs of Islam, How to Convert to Islam, Worship and Practice, The HEREAFTER, Stories of New Muslims, Comparative Religion, The Holy Quran, The Prophet Muhammad, Current Issues, Islamic History, Systems in Islam, Social Interaction, and Increasing Faith. The Contribution of each site in these categories are shown in figure 2. Table 1 shows a sample example of categories-topics articles in EIAD dataset. "The Evidence Islam is TRUTH" is a sample category which includes 178 articles distributed on 7 topics: The scientific miracles of the Holy Quran *with 15 articles, The scientific miracles of the prophet* Muhammed sayings with 2 articles, Muhammed in the Bible and other scriptures with 4 articles, The authenticity and preservation of the Holy Quran with 21 articles, Evidence of Muhammad's Prophethood with 29 articles, Logical proofs with 45 articles and The existence of God with 62 topics. Also, figure 3 shows a sample article with the primitive format. The article consists of two .txt files: content file and metadata file. Content file has the article content while the metadata file has the article name(*title), article description, number of topics if there, the number of views and the date of publishing.*

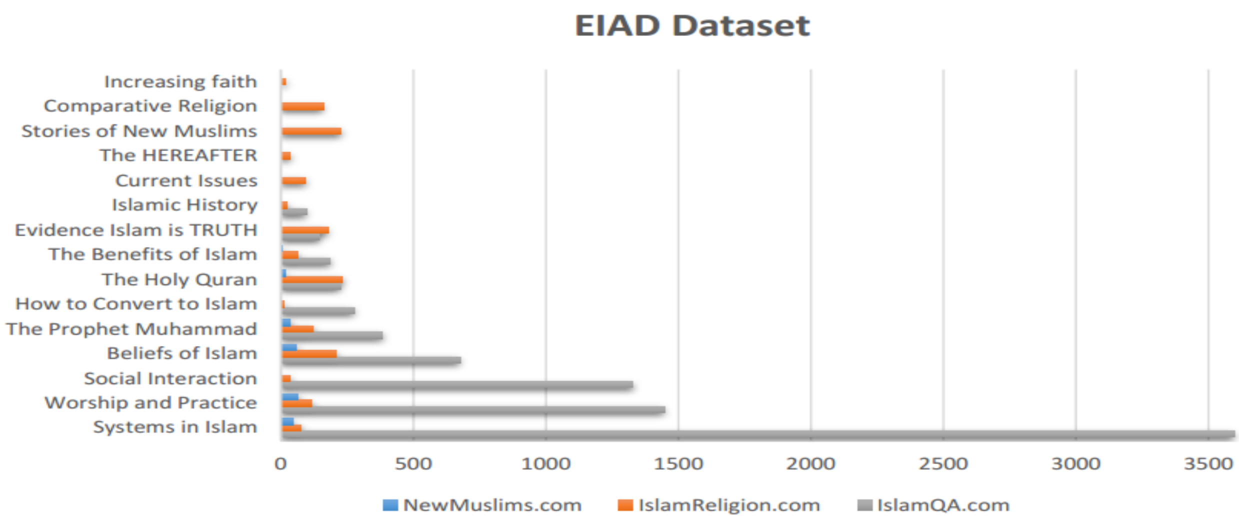


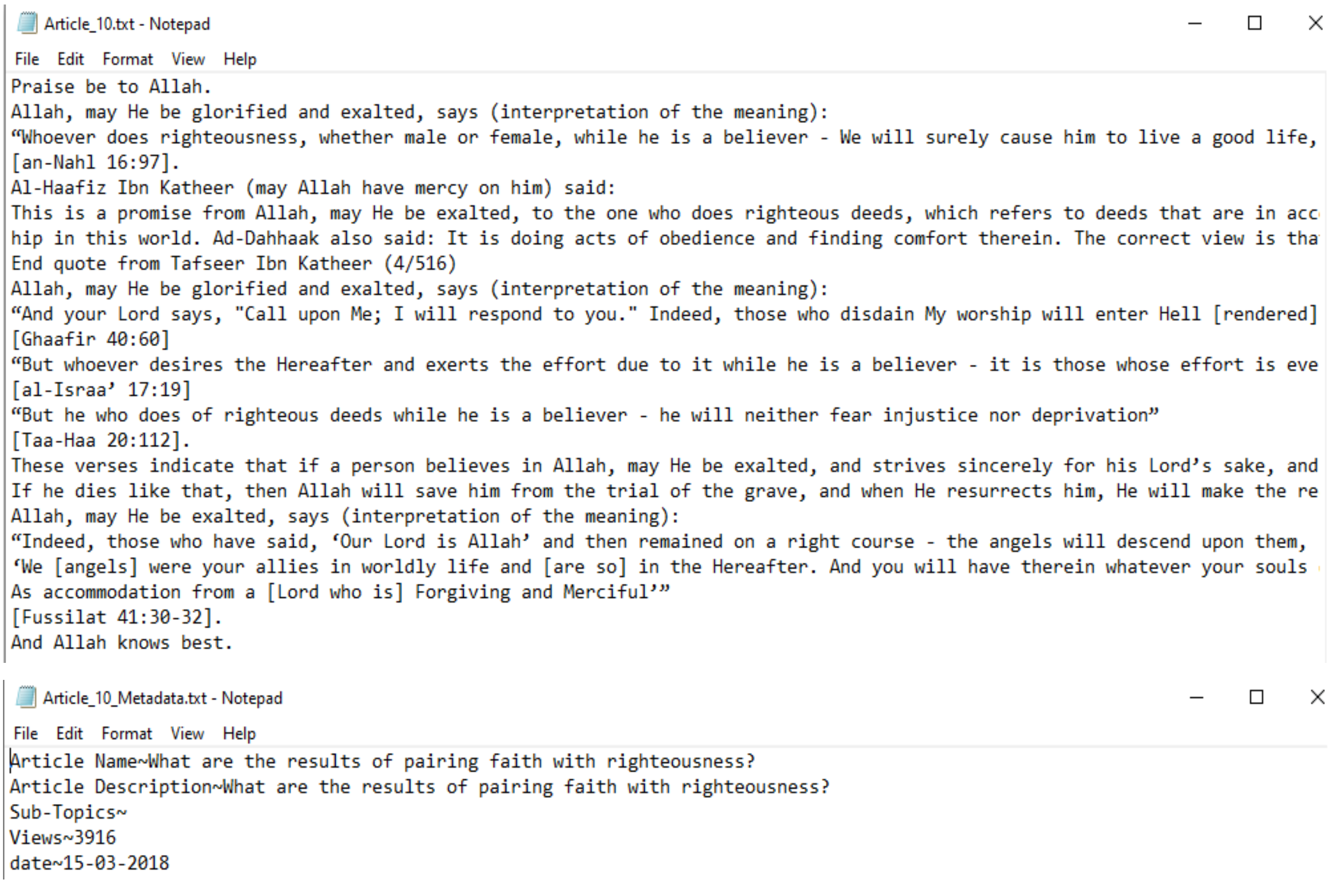

Figure 2 EIAD Dataset websites categories articles visualization

Once the dataset has been collected, it has to be cleaned. In data pre-processing step, the articles of EIAD dataset are cleaned. Data cleaning step includes some operations. Sentence segmentation is a process of splitting the articles into sentences. Then making tokenization about excruciating each sentence into separate words. The Following is removing the redundant unnecessary words from the articles like stop words, blank lines, pronouns, and prepositions. Finally, we apply text normalization to reduce the difference format of each word such as verbs may be in past tense, present tense, or future tense. Also nouns that may be singular or plural. We use the lemmatization algorithm to perform the text normalization. The lemmatization is responsible for getting the root of the words.

Table 1 Category-topics articles distribution

| Category | #Articles/ Category | Topics | #Articles/Topic |
|---|---|---|---|
| Evidence Islam is TRUTH | 178 | The Scientific Miracles of the Holy Quran | 15 |
| | | The Scientific Miracles of the Prophet Muhammad Sayings | 2 |
| | | Muhammad in the Bible and Other Scriptures | 4 |
| | | The Authenticity and Preservation of the Holy Quran | 21 |
| | | Evidence of Muhammad's Prophethood | 29 |
| | | Logical Proofs | 45 |
| | | The Existence of God | 62 |

### *3.1.2. Data Preprocessing*



Article_10.txt - Notepad
File Edit Format View Help
Praise be to Allah.
Allah, may He be glorified and exalted, says (interpretation of the meaning):
"Whoever does righteousness, whether male or female, while he is a believer - We will surely cause him to live a good life,
[an-Nahl 16:97].
Al-Haafiz Ibn Katheer (may Allah have mercy on him) said:
This is a promise from Allah, may He be exalted, to the one who does righteous deeds, which refers to deeds that are in acc
hip in this world. Ad-Dahhaak also said: It is doing acts of obedience and finding comfort therein. The correct view is tha
End quote from Tafseer Ibn Katheer (4/516)
Allah, may He be glorified and exalted, says (interpretation of the meaning):
"And your Lord says, "Call upon Me; I will respond to you." Indeed, those who disdain My worship will enter Hell [rendered]
[Ghaafir 40:60]
"But whoever desires the Hereafter and exerts the effort due to it while he is a believer - it is those whose effort is eve
[al-Israa' 17:19]
"But he who does of righteous deeds while he is a believer - he will neither fear injustice nor deprivation"
[Taa-Haa 20:112].
These verses indicate that if a person believes in Allah, may He be exalted, and strives sincerely for his Lord's sake, and
If he dies like that, then Allah will save him from the trial of the grave, and when He resurrects him, He will make the re
Allah, may He be exalted, says (interpretation of the meaning):
"Indeed, those who have said, 'Our Lord is Allah' and then remained on a right course - the angels will descend upon them,
'We [angels] were your allies in worldly life and [are so] in the Hereafter. And you will have therein whatever your souls
As accommodation from a [Lord who is] Forgiving and Merciful'"
[Fussilat 41:30-32].
And Allah knows best.

Article_10_Metadata.txt - Notepad
File Edit Format View Help
Article Name~What are the results of pairing faith with righteousness?
Article Description~What are the results of pairing faith with righteousness?
Sub-Topics~
Views~3916
date~15-03-2018

Figure 3 Primitive EIAD article format

### 3.1.3. *Data Preparing*

In spite of the EIAD dataset had been become in its cleaned form, we needed question-answers pairs beside these articles. Because our system architecture targets the long form question answering form, there wasn't models for generating questions with long answers. So, we had made crowdsourcing about specialized people in Islam religion with English language. This process was performed about using Haystack Annotation Tool [13]. Table 2 shows the actual annotated articles from EIAD dataset. It has the total number of articles in the dataset. Furthermore, the actual number of annotated articles using the Haystack annotation tool is around 7.5K. These annotated articles were used in generation of 10k question-answer pairs. All questions in the EIAD dataset have answers. Each question has at least one answer. These articles with their questions and answers were formulated in SQUAD dataset [14] format which is the most common format in the question answering task. Figure 4 shows sample of EIAD dataset in SQUAD format.

Table 2 EIAD Dataset Annotation size

| Articles | Annotated Articles | Question-Answer pairs |
|---|---|---|
| 10k | 7.5k | 10k |


```
{
  "paragraphs": [
    {
      "qas": [
        {
          "question": "Does the hadeeth, “Whoever among you sees an evil action, let him change it by his hand [by t
          "id": 343229,
          "answers": [
            {
              "answer_id": 365217,
              "document_id": 552773,
              "question_id": 343229,
              "text": "Changing evil things is to be done by stages, progressing by degrees from warning and remindi
              "answer_start": 44,
              "answer_end": 347,
              "answer_category": null
            }
          ],
          "is_impossible": false
        }
      ],
      "context": "Article_77_Metadata.txt\nPraise be to Allah.\nChanging evil things is to be done by stages, progre
      "document_id": 552773
    }
  ]
},
```


Figure 4 EIAD Dataset in SQUAD Format

### 3.1.4. *Data Vectorization*

As we know that any data has to be converted into vector representation. This vector representation facilitates any computations during training the computing module. Also, it simplifies the computations during the run time for the asked question. The dimensions of these vectors vary based on the used encoding models in the dense passage retriever models. Because of the difference in the lengths of the documents, there has to a padding operation to unify the length of all embedding vectors for these documents.

### 3.1.5. *Data Storing*

In the data storing component, we are ready to store our data including the dataset and the embedding of the documents of this dataset. The data store can be any database store like SQL server, Microsoft azure, or any local server. These stored documents are indexed for facilitating the searching process during the run time. We use SQL database with one of the most efficient indexing models which is called FAISS model [15] to store the EIAD dataset.

## 3.2. Computing Module

This is the second module in the NLP layer and the core module in the system architecture. It includes the main components of the Question Answering QA system (Retriever-Reader). These components are shown in figure 2. In the following sections, we will discuss each of them in details.

### *3.2.1. Retriever*

The retriever is the first component in the computing module. It is responsible for retrieving the relevant documents for the asked question. The used retriever is the Dense Passage Retriever DPR[16] which is sentence transformer-based. In most cases, DPR models outperform traditional retrievers (sparse retrievers) such as BM25 retrievers [17] since they are learnable, while traditional retrievers like BM25 rely on un-learnable techniques such as Bag of Words (BoW). As in figure 5 the DPR has two main encoders. The Encoder $E_D$ embeds all documents of the dataset. The Encoder$_q$ embeds the asked question $E_q$. At the run time, the dot product component computes the dot product and outputs the top k relevant documents.

### *3.2.2.Reader*

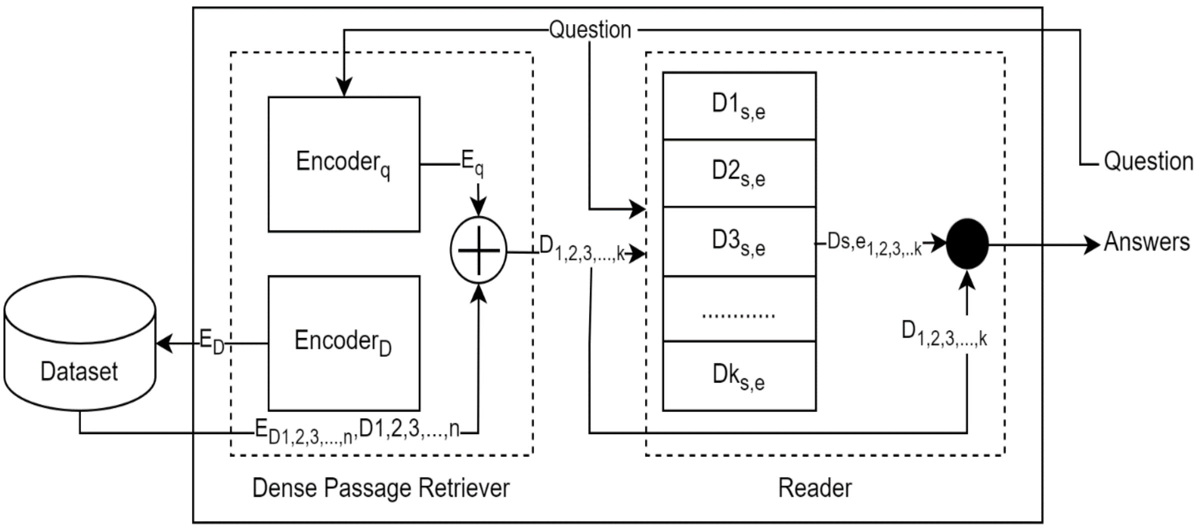



Figure 5: Computing Module

The second component in the computing module is the reader. The reader model is transformer-based model. This reader model takes the output of the DPR model (top k relevant documents) and a copy from the asked question q. It plays two main roles: generating the start s and the end e indexes of the answers in each top k documents then using these generated indexes with the top k documents to extracting the exact answers from the documents.

## 4. Case Study

This use case focuses on how to extract an answer for religious asked question. So, it assumes that the dataset is already collected and stored in database which means that the data module part is skipped during this case study. Also, the case study concentrates on discussing the process of the input\output I\O during the runtime as in figure 6 which are represented in the blue arrows. So, it explains the computing module which is responsible for taking the input (question) and providing the output (answer). The Retriever-Reader architecture in the case study depends on only the question as an input. the question q in the case study is **"What is the wisdom of fasting Ramadan?"** since that EIAD dataset is an Islamic religion dataset.

This question is encoded using fine-tuned Dense Passage Retriever DPR model $Encoder_q$. Then this $Encoder_q$ outputs the embedding vector of q as $E_{Question}$ in figure 7. Then the DPR retrieving the all documents (D1,D2,D3,……,Dn) in database with their stored embeddings ($E_{D1}$, $E_{D2}$, $E_{D3}$,……,$E_{Dn}$). The encoded question $E_{question}$ is compared with each of the retrieved document embeddings from database. This comparison is performed using the mathematical dot product then taking the highest top k results of these computations and output their equivalent relevant documents.

At this point, we have the top k relevant documents to the question from the database. A copy of the question with the top k relevant documents is passed into the reader model which working on extracting the starting s and the ending e positions of the answer from the relevant documents. Finally, it outputs the answers from these documents using the starting and the ending positions and showing the first one.

The inputs and outputs of these fine-tuned models are shown in the following figure in more detail. First of all, the question is passed into the Dense Passage Retriever DPR as an input as in figure 7. Then, the DPR is encoding this question and trying to get the most similar documents for this question encoding. The most similar documents are retrieved using the dot-product between the question encoding and the encodings of all database documents. the k value in this case study is 100. The top *100 documents related to the question are output from this retriever. This model achieves 67% Recall@100.* These most similar documents are passed into the fine-tuned reader model with a copy from the question as in figure 8. The reader takes the top 100 documents and output the answer of the asked question. This process is performed about extracting the start and the end indexes of the answer from each document. Then it shows the answers sorted from the most relevant one about the score in the last column in the table of figure 8. This model achieves F1 score 67.7%. After discussing a full case study of our ODQA system, Table 3 shows more examples on the input-output pairs of our ODQA system. These additional examples are shown from runtime test for the ODQA system. We can note the support of the system for the long-form answers .

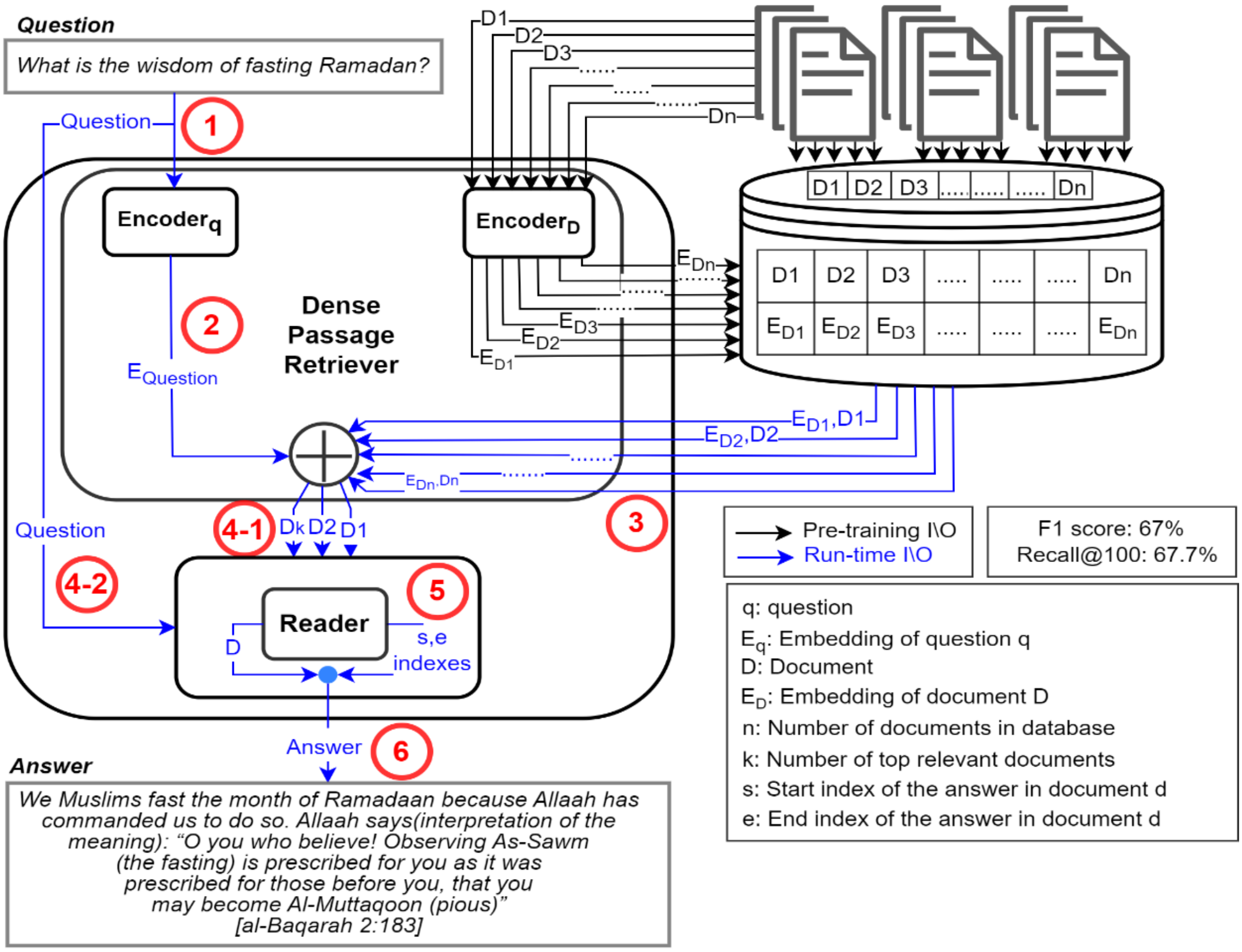
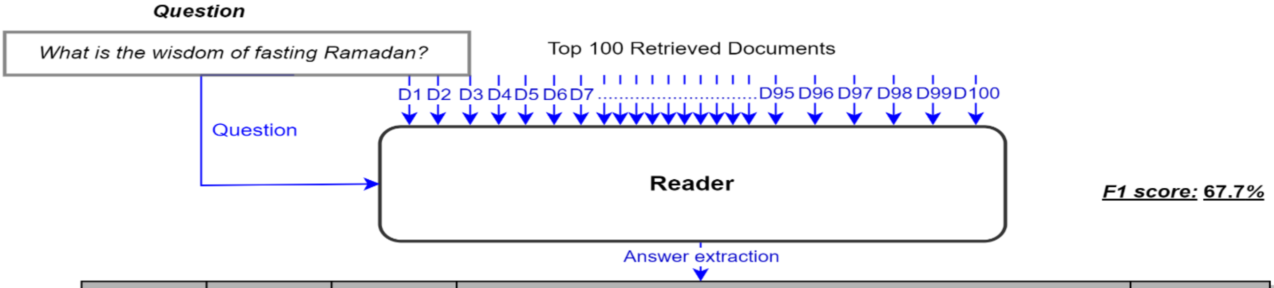


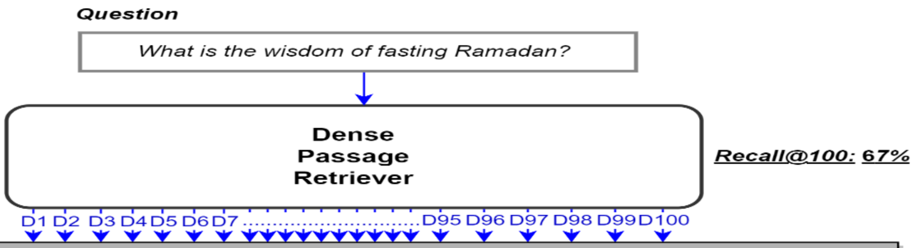

Figure 6: Retriever Reader Case Study

1. The asked question. 2. Asked question's embedding. 3. The embeddings of the documents in database with the content of these documents. 4-1. The top k retrieved documents related to the asked question. 4-2. Copy of the asked question. 5. Extraction of the answer indexes s (start) and e (end) from the retrieved documents. 6. Output the extracted answer.




| Doc. Num. | start index | end index | Answer | Score |
|---|---|---|---|---|
| 1 | 29 | 331 | We Muslims fast the month of Ramadaan because Allaah has commanded us to do so. Allaah says (interpretation of the meaning): “O you who believe! Observing As-Sawm (the fasting) is prescribed for you as it was prescribed for those before you, that you may become Al-Muttaqoon (pious)” [al-Baqarah 2:183] | 0.98 |
| 2 | 20 | 619 | The Ramadan fast is obligatory upon every accountable Muslim who is able to fast. If a Muslim is unable to fast because of a sickness that will harm him or cause great hardship for him if he does fast, or he needs treatment during the day in Ramadan in the form of pills, liquids and other things that are eaten or drunk, then in his case it is prescribed not to fast, because Allah, may He be exalted, says (interpretation of the meaning): “and whoever is ill or on a journey - then an equal number of other days. Allah intends for you ease and does not intend for you hardship” [al-Baqarah 2:185]. | 0.95 |
| .... | .... | .... | ........................................................................ | |
| 99 | 3395 | 3987 | It is not known that any of the salaf or early scholars said that it is prescribed to fast on any particular day of the week or the month or the year, and regard that day as a “festival” because the Prophet (blessings and peace of Allah be upon him) used to fast on the day of his birth every week, which was a Monday. If that was prescribed, the people of knowledge and virtue of the early generations who hastened to do all that is good would have hastened to do it before us. As they did not do that, it is known that this is something that is innovated and it is not permissible to do it. | 0.002 |
| 100 | 1531 | 1884 | You do not have to fast them one after another, rather it is permissible for you to fast them separately, according to what you are able to do. But you have to hasten to fast them and do not delay them again. You should start by making up the Ramadaan of the last year first, so that the next Ramadaan will not come before you have observed these fasts. | 0.002 |

Figure 7: Dense Passage Retriever Top 100 documents




| Document Number | Context |
|---|---|
| 1 | Praise be to Allah. Firstly: We Muslims fast the month of Ramadaan because Allaah has commanded us to do so. Allaah says (interpretation of the meaning): "O you who believe! Observing As-Sawm (the fasting) is prescribed for you as it was prescribed for those before you,that you may become Al-Muttaqoon (pious)" [al-Baqarah 2:183] So we worship Allaah by doing this act of worship which is beloved to Allaah and which He has enjoined upon us. The believers hasten to obey the commands of Allaah and His Messenger (peace and blessings of Allaah be upon him) ................................................................... ......................................................... .......................... |
| 2 | Praise be to Allah. The Ramadan fast is obligatory upon every accountable Muslim who is able to fast. If a Muslim is unable to fast because of a sickness that will harm him or cause great hardship for him if he does fast, or he needs treatment during the day in Ramadan in the form of pills, liquids and other things that are eaten or drunk, then in his case it is prescribed not to fast, because Allah, may He be exalted, says (interpretation of the meaning): "and whoever is ill or on a journey - then an equal number of other days. Allah intends for you ease and does not intend for you hardship" [al-Baqarah 2:185]. The Prophet (blessings and peace of Allah be upon him) said: "Allah loves for His concessions to be followed as He hates to be disobeyed. "Narrated by Imam Ahmad (5839); ................................. ............................ ........... |
| ....... | ................................... |
| 99 | Praise be to Allah. Did Prophet Muhammad fast on his birthday? Muslim (1162) narrated from Abu Qatadah al-Ansari (may Allah be pleased with him) that the Messenger of Allah (blessings and peace of Allah be upon him) was asked about fasting on Mondays and he said: "On (that day) I was born and on it Revelation came down to me." Al-Tirmidhi (747) narrated, in a hadith that he classed as hasan, from Abu Hurayrah (may Allah be pleased with him) that the Messenger of Allah (blessings and peace of Allah be upon him) said: "Deeds are presented (to Allah) on Monday and Thursday and I like my deeds to be presented when I am fasting." (Classed as saheh by al-Albani in Sahih al-Tirmidhi.) From the sahih hadiths quoted above, it is clear that just as the Prophet (blessings and peace of Allah be upon him) ............................................... ........................... ............... |
| 100 | Praise be to Allah. What the believer should do is to make the most of special occasions and worship his Lord a great deal at those times. One of the greatest of these occasions is the month of Ramadaan. How great are the acts worship that a Muslim does in Ramadaan, which bring a great reward. Ramadaan is the month of fasting, praying at night, reading Qur'aan, remembering Allaah (dhikr), making du'a', doing 'umrah, giving iftar to those who are fasting, giving charity, and being kind and generous to the needy. .............................................................. ...................................................... |

Figure 8: Reader Extracting Top 100 answers

Table 3 Examples on our ODQA system Question-Answers pairs

| Question | Answer |
|---|---|
| Why is it permissible for a man to marry more than one? | Marrying more than one wife is a matter which is recommended, subject to certain conditions: that the man be financially and physically able for it, and that he be able to treat his wives justly. Plural marriage brings a lot of benefits, including protecting the chastity of the women whom he marries, bringing people closer together, and producing a lot of children. This is what the Prophet (peace and blessings of Allaah be upon him) referred to when he said: “Marry the one who is loving and fertile”. And it serves a lot of other purposes. But for a man to marry more than one wife by way of boasting and taking that as a challenge to prove himself, this is a form of extravagance and extravagance is forbidden. |
| Is cheating forbidden in Islam? | Bribery Bribery means giving money to someone through whom a person takes something that he has no right to. For example, bribing a judge to get a verdict in your favor, or bribing an official to give you preference over others or to get other favors like allotment of a contract etc. |
| How is repentance in Islam? | You have to regret what you did, be determined not to repeat it, be sincere in your worship, do lots of naafil acts of worship such as praying naafil prayers night and day and observing voluntary fasts, read Qur’aan and make du’aa’. Allaah accepts the repentance of His slaves and forgives their bad deeds. |
| What are the rights of women in Islam? | Muslim women have the right to accept or refuse marriage proposals as they see fit, and married women are completely free from the obligation of supporting and maintaining the family. Workin g married women are free to contribute to the household expenses, or not, as they see fit. Women have the right to seek divorce if it becomes necessary. |

## 5. Conclusion

Open Domain Question Answering ODQA is a motivating field in NLP. During this work, we suggested a general retriever-reader based ODQA system architecture. It included two main layers: NLP layer and Interface layer which is platform independent. Also, this paper represented the used dataset (EIAD) which was created from scratch. EIAD dataset is an English Islamic QA dataset. The suggested system architecture has been discussed in details. Meanwhile, there was a real time case study has been shown for this ODQA system. This case study had only one input (question) then retrieved first 100 related documents for this question and extracted the top 100 answers from these documents. Finally, the system showed only the first related extracted answer. During this study, we attained reliable results using EIAD dataset which output scores: 67% Recall@100 for the DPR retrieval model and 67.7% F1 score for the reader model.

## References

[1] Weng; Lilian, Ed., “Open Domain Question Answering and Closed Domain Question Answering,” https://lilianweng.github.io/posts/2020-10-29-odqa/. 2020.

[2] Q. Zhang et al., “A Survey for Efficient Open Domain Question Answering,” ArXiv, vol. abs/2211.07886, 2022.

[3] Christiane Fellbaum, WordNet: An Electronic Lexical Database. Bradford Books, 1998.

[4] M. Mohammed, S. Amin, and M. M. Aref, "An English Islamic Articles Dataset (EIAD) for developing an IslamBot Question Answering Chatbot," in 2022 5th International Conference on Computing and Informatics (ICCI), 2022, pp. 303–309. doi: 10.1109/ICCI54321.2022.9756122.

[5] J. and Y. B. Zheng Dequan and Yang, "Open Domain Question Answering Based on Retriever-Reader Architecture," in Business Intelligence and Information Technology, 2022, pp. 723–733.

[6] P. S. H. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," CoRR, vol. abs/2005.11401, 2020, [Online]. Available: https://arxiv.org/abs/2005.11401

[7] D. S. Sachan, S. Reddy, W. L. Hamilton, C. Dyer, and D. Yogatama, "End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering," CoRR, vol. abs/2106.05346, 2021, [Online]. Available: https://arxiv.org/abs/2106.05346

[8] D. S. Sachan et al., "End-to-End Training of Neural Retrievers for Open-Domain Question Answering," CoRR, vol. abs/2101.00408, 2021, [Online]. Available: https://arxiv.org/abs/2101.00408

[9] K. Krishna, A. Roy, and M. Iyyer, "Hurdles to Progress in Long-form Question Answering," CoRR, vol. abs/2103.06332, 2021, [Online]. Available: https://arxiv.org/abs/2103.06332

[10] NewMuslims.com, Jan. 15, 2022.

[11] IslamReligion.com, Jan. 16, 2022.

[12] IslamQA.com, Jan. 25, 2022.

[13] "Haystack Annotation Tool," https://www.deepset.ai/annotation-tool-for-labeling-datasets, Jan. 30, 2022.

[14] "SQUAD - Stanford QUestion Answering Dataset," https://rajpurkar.github.io/SQuAD-explorer/.

[15] "FAISS," https://faiss.ai/, Dec. 30, 2022.

[16] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[17] G. Amati, "BM25," in Encyclopedia of Database Systems, M. T. LIU LING and ÖZSU, Ed. Boston, MA: Springer US, 2009, pp. 257–260. doi: 10.1007/978-0-387-39940-9_921.