End-to-End Open-Domain Question-Answering System: Baseline and Case Study using EIAD Dataset

Moataz Mohammed*, Salsabil A. El-Regaily, and Mostafa M. Aref
Ain Shams University, Abbassiya, Cairo, EGYPT

Abstract

During the artificial intelligence (AI) era, AI has evolved into a multidisciplinary industry in all domains. (NLP) Natural language processing, represents one of the most fascinating AI tasks. It can perform a number of tasks including question answering (QA), machine translation, entity linking, text generation, topic modelling, text summarization, and text to speech (TTS). The QA task is the focus of this research. It highlights the Open-Domain Question-Answering ODQA task explained using the field of Islamic religion. In this research, the QA task presents a model for developing an IslamBot QA system. IslamBot is a question-and-answer chatbot that is free-formed which can answer Islamic-related questions. The models of deep learning-based retrieval-reader were used to create the ODQA model. This paper uses a model that is based on data derived from the English Islamic Articles Database (EIAD). The EIAD dataset is a labelled ODQA dataset that was crowdsourced. The EIAD dataset contains approximately 10k articles, 7.5k of which were crowdsourced, and approximately 10k question-answer pairs. Every article contains at least one question-and-answer pair. This paper develops an end-to-end ODQA model that uses the EIAD dataset to create a benchmark and an entirely novel baseline model. It also sets a new standard with the most recent Dense Passage Retriever models, which achieve 78% R@100. The ODQA model also generated novel results. It received a 71.5% EM and a 75.8% F1 score. Furthermore, due to the length of the answer, the use of the long-form open-domain type is a hard issue: justification answer. Besides, the input of the model is only the question without context.

Key Words: Open-domain question answering; natural language processing; information retrieval; reading comprehension; retriever-reader.

1 1 Introduction

It has become indispensable to mimic human behavior during the past few decades. Computer Vision (CV) and Natural Answering (QA), Text Generation, Part of Speech (POS), Language Processing (NLP) are two subfields of Artificial Intelligence (AI) that mimic humans' vision and language. NLP is the subject of this study. A wide range of human language tasks can be performed using NLP. It is capable of performing many human language tasks such as: Text Summarization, Question Machine Translation, Named Entity Recognition (NER), and Text-to-Speech (TTS). The QA task is the focus of this research. It entails one person asking a question and another responding to it. It is possible to achieve this in the machine world by instructing the computer to mimic the responsible person for giving the answer [1]. The QA task can be done using NLP applications such as chatbots, which allow you to ask a question and receive an instant response from the developed chatbot. Chatbots can be created in a variety of fields, including economics, advertising, tourism, politics, ticket booking, social media, learning, call centers, industry, and religion. This study's use case is QA in the religious field. When looking for the possibility of using chatbots in Islamic websites, it was noticed that these websites are either knowledge-based or human-based. Human-based chatbots like Islam-Religion and Islam-Portal, are accessible twenty-four hours a day, seven days a week to answer any question. And as for the knowledge-based chatbots, it relies upon concepts such as knowledge graphs and decision trees, which can be found on websites such as Allah's Word, Islam House, Ask-A-Muslim, and Guide To Islam. Knowledge-based chatbots rely on a list of generic questions to select from until it finds the closest question to answer, but free-form chatbots are not available for these religious websites. Traditional QA systems are either closed-domain QA (CDQA) or reading comprehension (RC). In order to extract the answer from the RC systems, the user must provide the question and some context. However, progress proceeds in profound learning and the use of attention and processors. Systems for open-domain quality assurance (ODQA) has emerged. We can train a deep learning model on a large number of documents using open-domain QA systems. The model can then be completed by simply typing the question as input into the model without any context. This study makes a contribution by fine-tuning a long-form or free-form open-domain QA (ODQA) model on an Islamic religion dataset. The recent ODQA systems are either retriever-generator-

*Computer Science Department, Faculty of Computer and Information Sciences. Email: moataz.mohammed@cis.asu.edu.eg.

based or retriever-reader.

Furthermore, the data sets used by these systems can range from crowd-sourced datasets like Squad2 to datasets obtained from websites structured in a question-answer format, like the ELI5 dataset extracted from the Reddit website. LFODQA is a new QA sub-task, yet Hurdles [8] ranks among the most recent works of this kind of NLP task. With the ELI5 data set, it employs the retriever-generator model. Cluster Former Model [15] achieves cutting-edge performance in open domain question answering (ODQA) using a perfect match (EM) score of 68% and the Search QA dataset. With an EM score of 38.6%, the model of BERTserini [16] is an end-to-end open-domain QA model. Open-domain QA for COVID-19 [10] is a retriever-reader-based model with an EM score of 39.16% that uses Squad2 and COVID-QA datasets. We accomplished the following during this study:

1. A new benchmark in the open-domain question answering task using EIAD dataset.
2. A new end-to-end open-domain question answering model.
3. Cutting-edge results obtained while fine-tuning some of the most recent ODQA models and DPR models on the EIAD dataset.

Section 2 discusses related works on open-domain question answering as well as some research on QA datasets related to the Islamic domain. The end-to-end ODQA system architecture and the dataset used are then thoroughly discussed in Section 3. In Section 4 we demonstrate the results of our ODQA model experiments. The results are compared to the most recent state-of-the-art ODQA models. During the discussion Section 5, there are highlights for the results and models used followed by a case study in Section 6. Finally, during the conclusion in Section 7, a quick summary of this work is obtained.

2 Related Work

Before getting into our contributions to this study, we present major innovations in question answering tasks. This section discusses cutting-edge Open-Domain Question-Answering research. We concentrate on transformer-based research, such as[9, 13, 15], because they have made significant advances in Deep Learning in the last decade.

2.1 Learning Dense Representations of Phrases at Scale

The problem of answering open-domain questions can be reframed as a phrase retrieval issue. For the first time, we assert that we can learn dense representations of phrases on our own and reach much improved results in open domain QA. We present an efficient method [9] for learning phrase representations from reading comprehension tasks under supervision. We also recommend a query-side fine-tuning strategy to aid transfer learning and reduce the gap between inference and training.

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

Unsupervised pre-training with the Inverse Cloze Task and masked salient spans are followed by supervised fine-tuning using question-context pairs. This approach [13] leads to absolute gains of 2+ points over the previous best result in the top-20 retrieval accuracy on Natural Questions and TriviaQA datasets. We next explore two approaches for end-to-end training of the reader and retriever components in OpenQA models.

2.3 Cluster-Former: Clustering-based Sparse Transformer for Question Answering

Cluster-Former is a new sparse Transformer based on clustering that performs attention throughout chunked sequences. The proposed framework [15] is based on two distinct Transformer layers: the Cluster-Former Layer and the Sliding-Window one. This new design enables information integration beyond local windows, which is particularly useful for question answering (QA) tasks that depend on long-range dependencies.

3 Proposed System Architecture

The architecture of the proposed Open Domain Question Answering (ODQA) system is depicted in Figure 1. This system design is a retriever-reader paradigm that focuses on obtaining related articles and extracting the query response from these top-ranked articles. The EIAD dataset [11] was used to create this study. It is a collection of English Islamic articles. Crowdworkers used the Haystack annotation tool [4] to annotate this dataset. Each module of this design will be discussed in the parts that follow.

3.1 Database

Before moving on to the system's main components, we must first discuss the dataset that was used. The Content Table and the Indexing Table are the two main tables in the database. The Indexing Table tracks and stores the content embeddings. The Content Table displays the data from the used dataset. A Collection of English Islamic Articles During this work EIAD is the target dataset. This data set was gathered from three of the most reputable and secure Islamic websites on Internet, including IslamQA [5], Islam Religion [6], and New Muslims [12]. SQUAD is the format of the EIAD dataset. The dataset [11] appears to contain 10,000 articles divided into 15 Islamic categories. Each article has its own metadata, which includes the article title, description, rating, number of views, and date.

These articles were indexed using the FAISS index model [3] and stored in a SQL database. The overall number of articles in the dataset is shown in Table 1. Furthermore, the Haystack annotation tool has annotated approximately 7.5K articles. These annotated articles were used in question-answer pairs of generation 10k. The EIAD dataset contains answers to all of the questions. There is at least one answer to each question. Figure 2 shows that the length of these responses ranges from 50 to 1400 characters.

Figure 3 depicts the dataset's distribution. The EIAD dataset is divided into three subsets based on the most common distribution: training 80% with 6k annotated articles, then development 10% with 750 annotated articles, and testing 10% with another 750 annotated articles.

3.2 Retriever

Because the ODQA system is only concerned with the question, we must find the most relevant articles for this question in order to get the answer from such retrieved articles. A retrieving process is the process of finding the most relevant articles. The retriever element in Figure 1 is the ODQA system's main module. Dense Passage Retriever DPR-based [7] is the retriever here. It is equipped with a pair of encoders transformer-based [14] that serves as encoder models $E_{articles}$ and Equation. It also includes the FAISS model [3] for searching and the indexing tasks.

3.2.1 Encoder $E_{articles}$. In the dense passage retriever DPR models, Article Encoder $E_{articles}$ is the first model and component. This model functions as an encoder, taking in a text and producing a low-dimensional numerical representation vector for that text. Because this retriever model is a DPR model, the dataset that must be trainable on this DPR model must be in DPR format, which differs from the default SQUAD format. As a result, the EIAD dataset has a DPR-format replicated version. As a result, $E_{articles}$ is encoding all of the articles in the database. These encodings are then saved to be used later. They are used to select the query's most comparable K articles in the database with similar embeddings. We have more than one training trial for this model, which will be discussed further below.

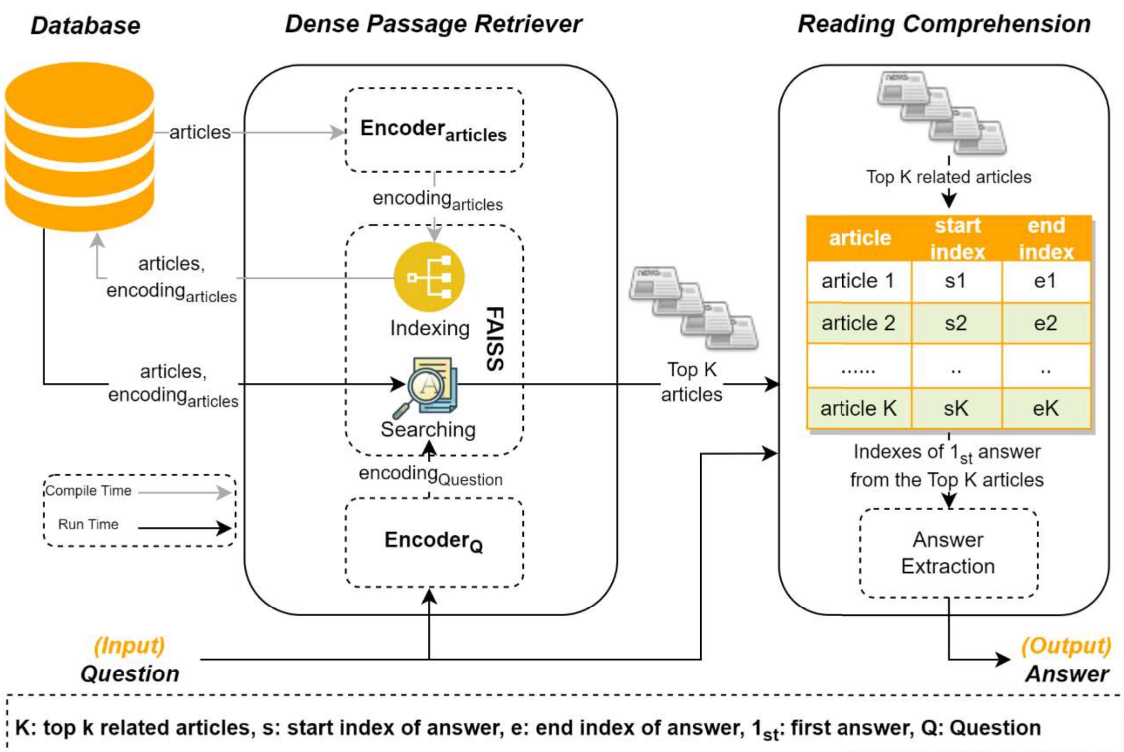


Figure 1: Open-domain question answering system architecture

Table 1: Dataset size

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

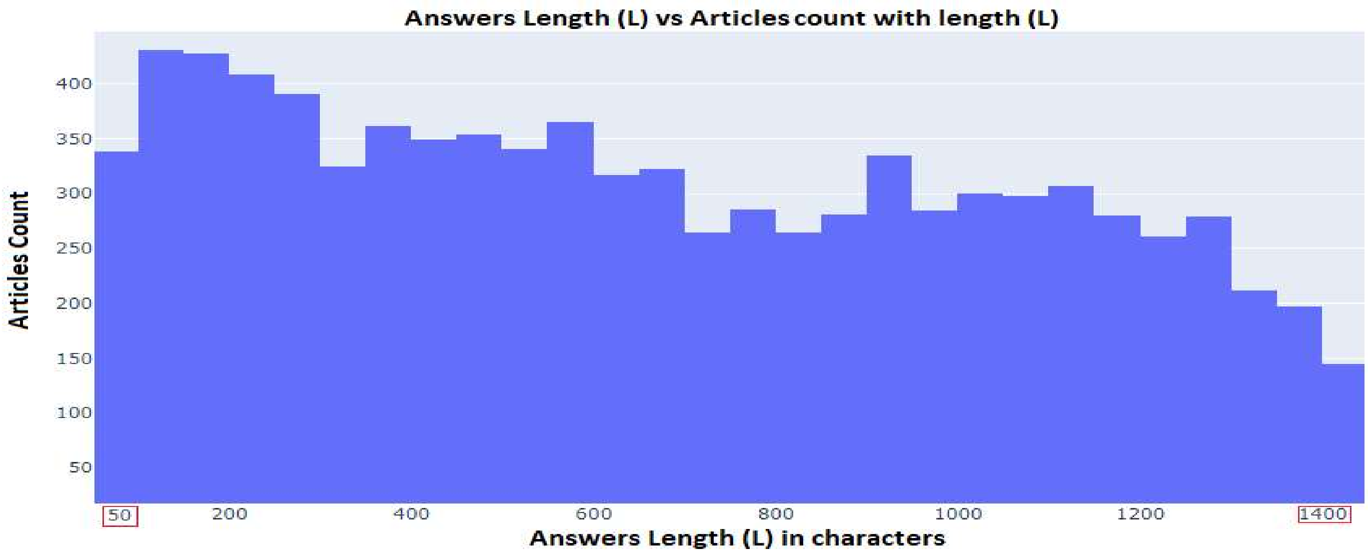


Figure 2: Articles count vs the range of answers length in characters

Figure 3: Dataset subset percentage

3.2.2 Encoder E_q. The articles encoder has the same concept. The question encoder model is an Eq model that encodes the asked question by taking the asked question as input text and applying the encoding process to it. As a result, it generates a low-dimensional vector for this question. This vector has a variable length and a maximum length of 50. The embedding of this question, along with the encodings of all articles, will be passed into the FAISS model later to extract the most similar articles. It is important to note that the previous models ($E_{articles}$ and E_q) are inter-correlated, which means that both had to be trained at the same time.

3.2.3 FAISS Model. Facebook created the FAISS model [3], which is a library. It is used to carry out an efficient search. The indexing procedure is what allows it to perform well in similarity searching. The indexing in the FAISS model is based on dense vector clustering. The FAISS model includes GPU and CPU support. In the following sections, we will go over the searching and indexing processes in more detail.

The FAISS flat index factory type was used to index the EIAD article dataset. The EIAD article dataset embeddings are stored with their text, like ($article_1$, $encoding_{article1}$), as a data structure of pairs, where article1 refers to the context of the first article and $encoding_{article1}$ refers to the content of the first article's encoding. So, the retriever takes the user's question and the encoder Eq encodes it in a dense vector at runtime. In addition, the retriever pulls all encodings from all the articles. The following formula is used to find the top maximum results by taking the dot product of the question vector $encoding_q$ and the vectors' articles $encoding_{articles}$. The indexes of the vector articles $encoding_{articles}$ that produced the most extreme results can then be used to obtain the top relevant articles. These indices are used to extract some other parts of pairs that include the article context.

3.3 Reader (Reading Comprehension)

The reader model's task is to extract the question's answer from the top-retrieved articles k. This was accomplished by learning how to extract the beginning and ending indexes of the answers from the original ones. Our reader is based on the Framework for Adapting Representation Models (FARM) reader. It is simple, quick, and easy to use. These readers are transformer-based, particularly the BERT [2] families. The FARM reader includes a prediction head and a built-in language model. In general, the reader is either an abstractive or an extractive reader. We use the extractive reader in our work because that domain is much more sensitive, requiring that the answer be extracted as it is.

$$articles_K = Top_K\left(DESC\left(\overrightarrow{embed_q}_{\ 1*M} \cdot \begin{bmatrix} \overrightarrow{embed_{art_1}} \\ \overrightarrow{embed_{art_2}} \\ \overrightarrow{embed_{art_3}} \\ \dots \\ \dots \\ \dots \\ \overrightarrow{embed_{art_N}} \end{bmatrix}^T_{N*M} \right) \right) \tag{1}$$

N: Dataset size

M: Embedding length

K: The number of articles to retrieve

4 Experimental Results

The dense passage retriever is the focus of this work when building the ODQA system. The answer was extracted from the retrieved article using the reader concept. The different trials of this system's components will be discussed during this section.

During these trials, more than one model is fine-tuned. Detailed information about each model and its results can be found in the following tables. Tables 2 and 3 show the results of the reader and retriever modules, respectively. Our own EIAD dataset was used for these experiments.

The DPR model was used to begin the training, and base-bert-uncased was used to encode the question and the article. However, the result was disappointing, reporting 33% recall@100. The recall@100 improved to hit 67% in the second trial while using Facebook's context encoder and question encoder models. All-MiniLM-L6-V2 is a sentence transformer-based boosted the recall@100 to 78% which exceeded all of the prior trials. The final attempt had the greatest outcome. The configurations of these several trials are shown in Table 2, including the number of parameters in each model as well as hyperparameters such as learning rate (LR), batch size, number of epochs, and embedding dimension. Table 2 also displays the earlier outcome of these improved models. Table 3 illustrates the various experiments for the reader model. With the exception of the metrics results and model dimensions, it displays the same entries from the retriever model's table. The reader models are known as QA models that respond to questions. So, we employed the F1 and exact match EM scores as measures to evaluate the QA models. Two models—distilled-bert-base-uncased-distilled-squad and Roberta-base-squad2—were used for the majority of the trials. Based on the change in batch size and the number of epochs, the first model obtained two distinct scores for the F1 score and EM score. Using 10 batch sizes and 8 epochs, it reached 65.57% F1 score and 59.33% EM. With 16 epochs, it attained 75.8% F1 score and 71.5% EM. The second variation, Roberta-base-squad2, scored (67.7%, 67.4%) for F 1 scores and (60.5%, 61.4%) for EM. The distilled-bert-base-uncased-distilled-squad achieved the greatest results after all of these tests, with a 75.8% F1 score and 71.5% EM.

Table 2: DPR model trials

Model		**Results**	**Batch Size**	**Epochs**	**Learning Rate**	**Dimension**	**Number Parameters**	
question	**context**	**R@100**					**question**	**context**
bert-base-uncased	bert-base-uncased	33%	16	16	3e-05	768	110M	110M
facebook/dpr-question-encoder-single-nq-base	facebook/dpr-ctx-encoder-single-nq-base	67%	8	16	3e-05	768	110M	220M
sentence-transformers/all-MiniLM-L6-v2	sentence-transformers/all-MiniLM-L6-v2	**78%**	10	16	3e-05	384	23M	23M

Table 3: Reader models trials

Model	Results		Batch Size	Epochs	Learning Rate	Number Parameters
Name	Exact match (EM)	F1				
distilbert-base-uncased-distilled-squad	59.33%	65.57%	10	8	1e-05	66M
roberta-base-squad2	60.5%	67.7%	10	8	1e-05	125M
roberta-base-squad2	61.4%	67.4%	32	16	1e-05	125M
distilbert-base-uncased-distilled-squad	**71.5%**	**75.8%**	10	16	1e-05	66M

Once we reached the best outcomes for the ODQA components on EIAD dataset, we can perform comparison with some of recent works in ODQA task. Comparison between our best DPR results and one of the best models in this task and the inventor of the DPR concept [7] is shown in Table 4. This comparison is constructed by fine-tuning the original DPR model on our EIAD dataset. Table 4 shows that our best DPR component trial beats the original DPR model by increasing the retrieval accuracy on the Top-100 and Top-1 by 11%. Furthermore, we contrasted one of the most current ODQA works with our Reader model (ODQA model) [16] which was fine-tuned on EIAD dataset. The EM and F1 score of this model are distinguished with those of our model in Table 5. Table 5 shows that our model performs better than [16] by 43% EM and F1 score.

5 Discussion

We have discussed an end-to-end open domain question answering ODQA task during this research. In addition, by providing the asked question without context, we demonstrate how ODQA provides a more exciting task than traditional QA. We focused on how to make an effective end-to-end ODQA system in this study, as efficiency is achieved by optimizing storage resources and GPUs used. The storage resources were optimized through one of the following ways: 1. To avoid zero values, use the Dense Passage Retriever DPR models rather than the sparse retrieval models. 2. Despite the fact that the DPR makes use of used storage, it still employs two models. As a result, we sought models that were as light as possible while maintaining accuracy. During the DPR, these models were used to compensate the model's weights of the complex ones while preserving storage optimization. The same was true for the reader model, which attempted to select a small model with excellent accuracy. GPU optimization was performed using the maximum number of batch sizes allowed for the model during fine-tuning. The maximum number of batch sizes varied from model type to model type based on model size. In spite of all this, there are some limitations with this architecture. The dimension of the retriever input

Table 4: Comparison table based on Top-1 & Top-100 retrieval accuracy of models fine-tuned on EIAD dataset

Model	Top-1	Top-100
Dense Passage Retrieval for Open-Domain Question Answering [14]	13.5	67.5
all-MiniLM-L6-V2-distilbert-base-uncased-distilled-squad (ours)	**24.4**	**78.8**

Table 5: Comparison table based on EM & F1 scores for reader models fine-tuned on EIAD dataset

Model	EM	F1
End-to-End Open-Domain Question Answering with BERTserini [4]	28.23	41.36
all-MiniLM-L6-V2-distilbert-base-uncased-distilled-squad (ours)	**71.5**	**75.8**

(question) vectors and context vectors were limited to 384 which achieved the best result. Also, the dataset must be annotated to train the model. Additionally, due to the small size of the dataset, there was an accuracy limit. The EIAD dataset contained approximately 10,000 annotated question-answer pairs. If the number of annotated question-answer pairs increased, the model's accuracy might get improved as more features were learned.

6 Case Study

This case study focuses on obtaining an answer to a religiously stated question on the Top-500 retrieved articles. As a result, it presupposes that the dataset has already been stored and collected in the database, implying that the database module portion of this case study is omitted. Furthermore, the case study focuses on exploring the process of input – output IO throughout runtime, as seen in Figure 4, which is represented by blue arrows. As a result, it illustrates the question answering system that is in charge of receiving input (query) and producing output (answer). In the example study, the Retriever-Reader architecture relies solely on the query as an input. The case study's question q is **"Why there are heaven and hell?"** because the EIAD dataset is an Islamic religion one. $Encoder_q$, a fine-tuned Dense Passage Retriever DPR model, is used to encode this query. Then, as seen in Figure 5, $Encoder_q$ outputs the embedding vector of q as $E_{question}$. The DPR then retrieves all documents (D_1,D_2,D_3,......,D_n) in the database along with their associated embeddings (E_{D1}, E_{D2}, E_{D3},......,E_{Dn}). The embedded query $E_{question}$ is compared to each of the database-retrieved document embeddings. This comparison is carried out by computing the mathematical dot product and then taking the top k outcomes of these computations and producing their comparable relevant articles. At this point, we have the top k relevant articles from the database to the query. A copy of the question along with the top k relevant papers is fed into the reader model, which works on deriving the starting s and ending e locations of the response from the relevant articles. Lastly, it displays the first of the responses from these papers based on the starting and ending places.

The inputs as well as the outputs of such fine-tuned models are depicted in further detail in the following image. The Dense Passage Retriever DPR receives the query as an input, as shown in Figure 5. The DPR then encodes this question and attempts to find the most related articles for this question encoding. The dot-product of the encodings of all database articles and the question encoding is used to retrieve the most relevant articles. In this case study, the k value is 500. This retriever returns the top 500 articles linked to the question. This model obtains a Recall@500 of 87.8%.

As seen in Figure 6, our reader model takes the most similar articles as input together with a replica from the query. The reader selects the top 500 articles and outputs the response of the query.

This procedure involves retrieving the beginning and ending positions of the response from each article. Then, in the last column of Figure 6, it displays the responses ordered by the most relevant one concerning the score. This model has an F1 score of 75.8%.

7 Conclusion

Throughout this research, we explained the QA task in NLP and focused on the ODQA models, which were the intended task. In addition, recent studies related to this NLP task have been displayed. Following that, we demonstrated our ODQA system architecture, which included our EIAD dataset. This system architecture was based on a retriever-reader model, with two main models: reader and retriever. Each of these models had more than one trial. All of these trials were fine-tuned using the EIAD dataset. The most difficult challenge during this project was managing its resources. All trials were run on either a single RTX 2080ti GPU with 12GB VRAM or a single GTX 1080ti GPU with 12GB VRAM. Because DPR models contain two models internally as it is heavy, the RTX 2080ti was used to train them. The GTX 1080ti was used in conjunction with lighter models i.e., reader models. Another difficulty was the large size of the dataset. Our dataset (EIAD dataset) was 10k question-answer pairs in size, compared to the SQUAD dataset, which was 100k, and SUQAD2, which was 130k. Although we achieved new results for these models, the DPR model outperformed the reference model [7], which reached 78.8% top-100 accuracy with all MiniLM-L6-V2 and 78% recall@100 score while the reference model [7] reported 67.5% top-100 accuracy and 67% recall@100. Similarly, the reader model achieved 75.8% F1 score and 71.5% EM by using distilbert base-uncased-distilled-squad, which is the best outcome compared to [16], which reported 41.36% F1 score and 28.23% EM. As a result, we created a new baseline for the EIAD dataset called the all-MiniLM-L6-V2-base-uncased-distilled-squad-EIAD.

References

[1] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," *Procedia Comput Sci*, 73:366–375, 2015, doi: https://doi.org/10.1016/j.procs.2015.12.005.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018.

[3] FAISS 2022, Accessed 30 December 2022, https://faiss.ai/.

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

[5] "Islam Question and Answer 2022," Access Jan. 25, 2022, https://islamqa.info/en.

[6] "Islam Religion 2022," Accessed Jan. 16, 2022, https://www.islamreligion.com/.

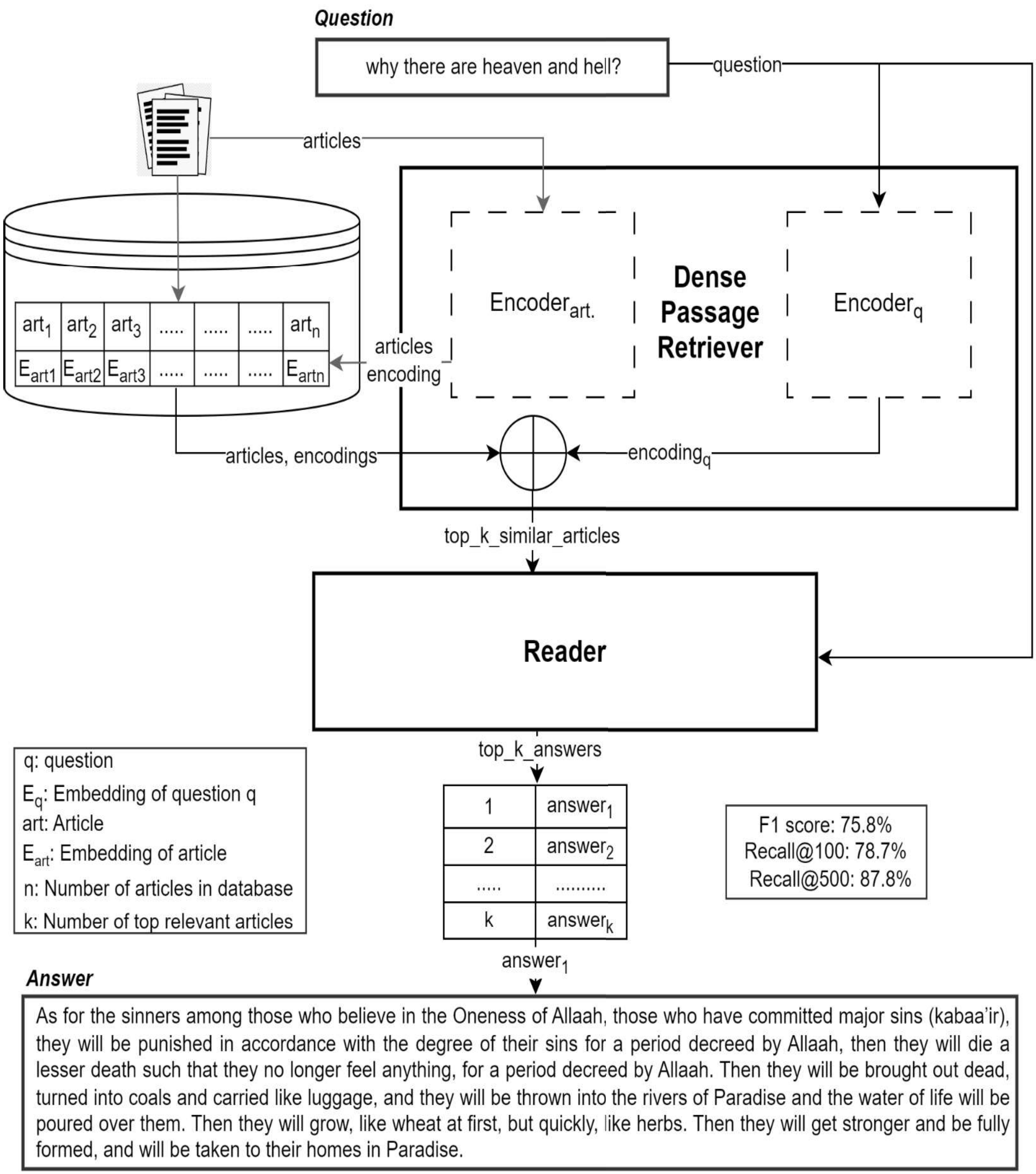
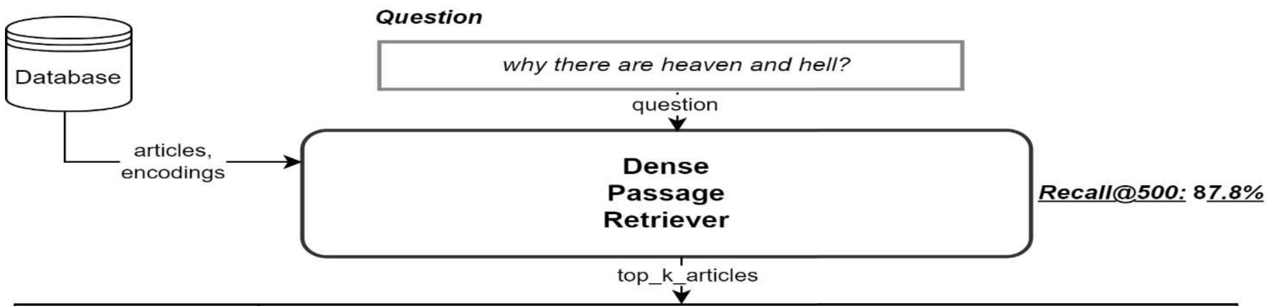
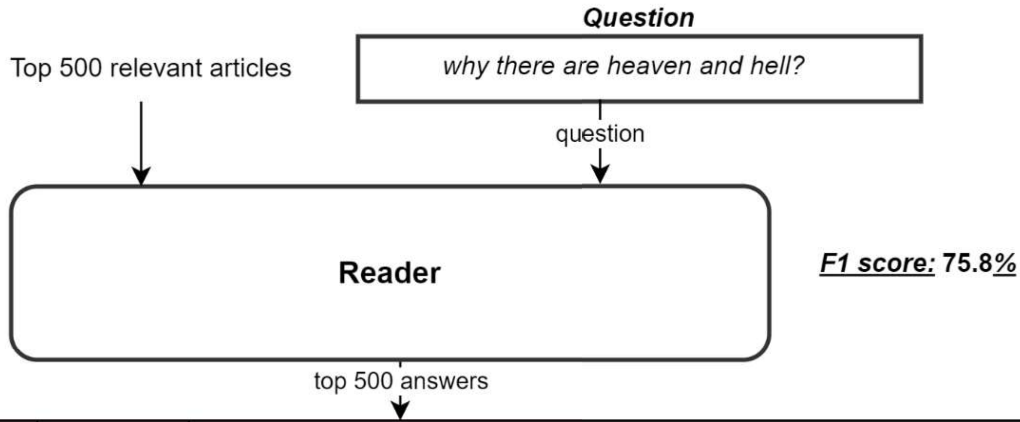


Figure 4: Retriever reader case study



article index	content
1	Praise be to Allah. We should understand properly the general principle concerning this matter, the matter of entering Paradise and spending eternity in Hell. It is a simple matter that is explained in a brief hadeeth that was narrated by Muslim in his Saheeh (135) from Jaabir (may Allah be pleased with him) who said: A man came to the Prophet (blessings and peace of Allah be upon him) and said: O Messenger of Allah, what are the two deeds that make entering Paradise or Hell inevitable? He said: "Whoever dies not associating anything with Allah will enter Paradise, and whoever dies associating anything with Allah will enter Hell." An-Nawawi said: With regard to the words, "What are the two deeds that make entering Paradise or Hell inevitable?" what is meant are the characteristic that makes Paradise inevitable and the characteristic that makes Hell inevitable. End quote. This hadeeth explains that what makes it inevitable that a person will enter Paradise is if he dies believing in Tawheed.. ....
2	Praise be to Allah. Firstly: Paradise has degrees or levels (we ask Allah to make us among its people), and Hell also has degrees or levels (we seek refuge with Allah from it). The people of Paradise will vary in their degrees or levels, according to the level of their faith and righteous deeds in this world. The best of them in knowledge, righteous deeds and faith will be the highest of them in the levels of Paradise. The people in the lowest levels will not be able to attain what is in the highest levels, because they did not do that which makes them deserving of attaining those levels. If all the people of Paradise were to share in the bliss that Allah has prepared for those who are above them, then there would be no wisdom in the variation of status and degree! By Allah's perfect justice, those who are deserving of Paradise will not all be the same in degree or level of bliss. Variation between people in this world in terms of faith and obedience leads to variation in their status and standing before Him, may He be glorified and exalted. See the answer to question no. 126349 . Secondly: The people of Paradise will be in a state of eternal bliss, whether they are of the highest levels or less than that. There they will have whatever they wish for, as Allah, may He be exalted, says (interpretation of the meaning): "Gardens of perpetual residence, which they will enter, beneath which rivers flow.. ....
....	....
499	Praise be to Allah. Some people have started to claim that the Sunnah is not a source of legislation. They call themselves al-Quraaniyyeen and say that we have the Quraan, so we take as halaal whatever it allows and take as haraam whatever it forbids. The Sunnah, according to their claims, is full of fabricated ahaadeeth falsely attributed to the Messenger of Allaah (peace and blessings of Allaah be upon him). They are the successors of other people about whom the Messenger of Allaah (peace and blessings of Allaah be upon him) told us. Ahmad, Abu Dawood and al-Haakim reported with a saheeh isnaad from al-Miqdaam that the Messenger of Allaah (peace and blessings of Allaah be upon him) said: Soon there will be a time when a man will be reclining on his couch, narrating a hadeeth from me, and he will say, Between us and you is the Book of Allaah: what it says is halaal, we take as halaal, and what it says is haraam, we take as haraam. But listen! Whatever the Messenger of Allaah forbids is like what Allaah forbids. (Al-Fath al-Kabeer, 3/438. Al-Tirmidhi reported it with different wording... ....
500	Praise be to Allah. Firstly: Before replying to this question, we must establish an important point about the virtues of certain soorahs. There are fabricated ahaadeeth about the virtues of various soorahs which have been falsely attributed to the Messenger of Allaah (peace and blessings of Allaah be upon him). Among the most famous of those who are known for that are the following: 1 – Nooh ibn Abi Maryam al-Jaami', of whom it was said: He encompassed everything except the truth. He regarded it as permissible to tell lies in hadeeth in the interests of the religion, and he made up ahaadeeth by himself and attributed them to the Messenger (peace and blessings of Allaah be upon him) concerning the virtues of the soorahs of the Qur'aan, soorah by soorah.. ....

Figure 5: Dense passage retriever top 500 documents



article index	start index	end index	Answer
1	1275	1965	As for the sinners among those who believe in the Oneness of Allaah, those who have committed major sins (kabaa'ir), they will be punished in accordance with the degree of their sins for a period decreed by Allaah, then they will die a lesser death such that they no longer feel anything, for a period decreed by Allaah. Then they will be brought out dead, turned into coals and carried like luggage, and they will be thrown into the rivers of Paradise and the water of life will be poured over them. Then they will grow, like wheat at first, but quickly, like herbs. Then they will get stronger and be fully formed, and will be taken to their homes in Paradise.
2	20	401	The scholars of Ahl al-Sunnah wa'l-Jamaa'ah are agreed that Paradise and Hell are two created things that exist at present. None of them doubt that because of the volume of evidence from the Qur'aan and Sunnah which indicates that. From the Qur'aan: Allaah says (interpretation of the meanings): "[Paradise] prepared for Al-Muttaqoon (the pious)" [Aal 'Imraan 3:133]
....	....	....	....
499	4275	4705	if he attains the highest level in Paradise, he will not miss out on any blessing that he hopes to attain in any of the lower degrees, so there is no need to ask for the lower degrees. But it will be possible for him to go down to a lower degree in order to visit a brother of his there – as some of the scholars said – so that he may acknowledge the blessing that Allah has bestowed upon him and His great generosity.
500	2329	3093	Ibn Al-Qayyim (may Allah have mercy on him) said: That is because these people will enter Paradise without being called to account because of the perfection of their Tawheed, therefore he described them as people who did not ask others to perform ruqyah for them. Hence he said "and they put their trust in their Lord." Because of their complete trust in their Lord, their contentment with Him, their faith in Him, their being pleased with Him and their seeking their needs from Him, they do not ask people for anything, be it ruqyah or anything else, and they are not influenced by omens and superstitions that could prevent them from doing what they want to do, because superstition detracts from and weakens Tawheed.

Figure 6: Reader extracting top 500 answers

[7] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” *CoRR*, abs/2004.04906, 2020.

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

[9] J. Lee, M. Sung, J. Kang, and D. Chen, “Learning Dense Representations of Phrases at Scale,” *CoRR*, abs/2012.12624, 2020, Online: Available: https://arxiv.org/abs/2012.12624.

[10] S. Levy, K. Mo, W. Xiong, and W. Y. Wang, “Open-Domain Question-Answering for COVID-19 and Other Emergent Domains,” *CoRR*, abs/2110.06962, 2021, Online: Available: https://arxiv.org/abs/2110.06962.

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

[12] “New Muslims 2022,” Accessed Jan. 15, 2022, https://www.newmuslims.com/.

[13] D. S. Sachan, M. Patwary, M. Shoeybi, N. Kant, W. Ping, W . L. Hamilton, and B. Catanzaro, “End-to-End Training of Neural Retrievers for Open-Domain Question Answering,” *CoRR*, abs/2101.00408, 2021, Online: Available: https://arxiv.org/abs/2101.00408.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, J. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All You Need,” *CoRR*, abs/1706.03762, 2017, Online: Available: http://arxiv.org/abs/1706.03762.

[15] S. Wang, L. Zhou, Z. Gan, Y-C Chen, S. Sun, Y. Fang, Y. Chen, and J. Liu, “Cluster-Former: Clustering-based Sparse Transformer for Long-Range Dependency Encoding,” *ACL-IJCNLP 2021*, Aug. 2021.

[16] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, “End-to-End Open-Domain Question Answering with BERTserini,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 72–77, June 2019, doi: 10.18653/v1/N19-4013.

Moataz Mohammed is a teaching assistant of Computer Science Department at Ain Shams University, Cairo, Egypt. He is a Master candidate student in NLP field, B.Sc. of Computer Science Dept., Faculty of Computer and Information Sciences, in 2017, Ain Shams University, Cairo, Egypt.

Salsabil A. El-Regaily received her Ph.D. in 2019 and M.S. in 2013, both in computer science from University of Ain Shams, Cairo, Egypt. She also received a B.S. in 2007 in Computer Sciences from Ain Shams University, Cairo, Egypt. Her research interests are machine learning and medical image processing

Mostafa M. Aref is a Professor of Computer Science and Chairman of Computer Science Department, Ain Shams University, Cairo, Egypt. He received a Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. He obtained his M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask, Canada and a B.Sc. of Electrical Engineering - Computer and Automatic Control Section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.