Is GPT-4 a Good Islamic Expert for Answering Quran Questions?

Conference Paper · December 2023

1 author:

Sarah Alnefaie
King Abdulaziz University
12 PUBLICATIONS **41** CITATIONS

Is GPT-4 a Good Islamic Expert for Answering Quran Questions?

Sarah Alnefaie
King Abdulaziz University,
Jeddah, Saudi Arabia
University of Leeds, Leeds, UK
`scsaln@leeds.ac.uk`

Eric Atwell
University of Leeds
Leeds, UK
`e.s.atwell`
`@leeds.ac.uk`

Mohammad Ammar Alsalka
University of Leeds
Leeds, UK
`m.a.alsalka`
`@leeds.ac.uk`

Abstract

In this research, we investigated GPT-4 as a question-answering model for the Holy Quran. As a first step, we built the Quran question–answer pair (QUQA) dataset, comprising 2,189 questions, and made it freely available via our repository. This dataset was then used to benchmark the performance of the current Generative Pre-trained Transformer 4 (GPT-4) model from the OpenAI research laboratory. The results show that GPT-4 did not do well with this dataset, with a 0.23 partial Average Precision (pAP) score, 0.26 F1@1 score, and 0.19 Exact Match (EM) score. Therefore, further improvement is needed for Classical Arabic responses generated by GPT model.

Keywords: GPT-4, large language model, and Quran question–answer pair (QUQA) dataset

1 Introduction

Vaswani et al. (2017) introduced the transformer architecture that significantly advanced the field of natural language processing (NLP). Bidirectional encoder representations from transformers (BERT) is one of the pre-trained transformer language models that has outperformed the state of the art in many NLP downstream tasks, such as text classification, question answering and machine translation (Devlin et al., 2018).

A challenging task in NLP is answering questions in Arabic, especially from Classical Arabic texts, such as the Quran and Hadith Sharif. The challenge lies in the nature and features of the text. For example, the meaning and spelling of the terminology in the Quran differ from other Arabic variants (colloquial Arabic, Modern Standard Arabic and Classical Arabic) (Altammami and Atwell, 2022). This challenge has motivated several researchers to use the BERT transformer for a Holy Quran question-answering system, and it

has shown promising results (Ahmed et al., 2022; Alsaleh et al., 2022; ElKomy and Sarhan, 2022; Malhas and Elsayed, 2022)

Recently, OpenAI has developed and provided access to several versions of the large language model (LLM)-driven ChatGPT—"GPT" refers to its underlying generative pre-trained transformer model (Brown et al., 2020; Ouyang et al., 2022). Several studies have focused on testing GPT on downstream tasks (Jiao et al., 2023; Wang et al., 2023; Qin et al., 2023). However, to the best of our knowledge, there has not yet been a published examination of how well GPT works in answering Islamic questions.

Our research sought to answer the question, "Is GPT a good Islamic expert?" To answer this question, we first created a corpus for Islamic questions and answered them using the Noble Quran due to the lack of an available dataset of this type. We then tested the performance of GPT with these questions to learn about its behaviour, strengths and disadvantages. The results of this study will benefit a large segment of Muslims worldwide, motivate additional research to improve upon any defects found and help researchers choose an appropriate language model in the future.

2 Related Work

2.1 Islamic Question-and-Answer Systems

Many studies have built systems for finding answers to Islamic questions from the Holy Quran and Hadith Sharif. These studies typically used three different approaches.

2.1.1 Question-and-Answer Systems Based on Retrieval Techniques

Some research has focused on the retrieval techniques used to find answers from Quranic text. Abdelnasser et al. (2014) developed the Al-Bayan sys-

tem. This system receives the question as input and then retrieves a Quran verse that includes the answer based on ontology by computing the cosine semantic similarity between the question and the concept vectors. Their study found that the system had an accuracy of 85%. Abdi et al. (2020) proposed a model that answers questions from the Hadith by calculating the semantic similarity between the question and Hadith sentences utilising an Arabic dictionary. Maraoui et al. (2021) introduced a Hadith question-and-answer system based on two basic stages. First, more information is added to the question. Second, the Hadith text is represented using a Text Encoding Initiative (TEI) standard. The accuracy of the system was 92%. This approach can generally answer only factual questions.

2.1.2 Question-and-Answer Systems Based on Knowledge Bases

Hamoud and Atwell (2016) recommend building a simple search system over a close-domain knowledge base to answer all kinds of questions about the Quran. First, they built a corpus of 1,500 questions and their answers. The dataset included different types of questions. A simple matching process was applied to a user's query and the questions in the dataset to find the most relevant question and display its answer. This system demonstrated 79% precision and 76% recall. The drawback of a knowledge base approach is that a model's performance is entirely affected by the corpus's quality, including the variety and quantity of the information it contains.

2.1.3 Question-and-Answer Systems Using Pre-Trained Language Models

Many studies have used pre-trained language models for Quran Machine Reading Comprehension (MRC) task. The model inputs are typically a Modern Standard Arabic question and a Quranic passage in Classical Arabic (CA), while the output is a ranked list of five answers. Each answer is part of the passage. They have relied on the Qur'anic Reading Comprehension Dataset (QRCD) for the training and testing phases.

ElKomy and Sarhan (2022) developed a system by training five different Arabic models: Arabic BERT (ARBERT), AraBERTv02-Base, AraBERTv02-Large, masked Arabic BERT (MARBERT) and QCRI [Qatar Computing Research Institute] Arabic and Dialectical BERT (QARiB)-

Base. They used a voting system among these models to select the final answer. To improve the results, they applied post-processing steps. Their experiments showed the following results: 0.27 Exact Match (EM), 0.50 F1@1 and 0.57 partial Reciprocal Rank (pRR).

Ahmed et al. (2022) proposed augmenting the QRCD dataset manually with 657 questions and answers to train their Arabic Efficiently Learning an Encoder that Classifies Token Replacements Accurately (AraELECTRA) and the Arabic-Typologically Diverse Question Answering (ArTyDiQA) language model. This model was able to extract the answers to the questions well, achieving the following results: 0.24 EM, 0.51 F1@1 and 0.55 pRR.

Mostafa and Mohamed (2022) recommend using the AraELECTRA model after fine-tuning it on the Arabic-Stanford Question Answering Database (SQuAD), ArTyDiQA and Arabic Reading Comprehension Dataset (ARCD). Finally, they used different loss functions to address the data imbalance issue. The study found that the model achieved 0.55 pRR.

Malhas and Elsayed (2022) conducted an experiment by fine-tuning AraBERT and Classical Arabic BERT (CL-AraBERT), using the MSA datasets and the training portion of the QRCD and then assessing the models using the test questions from the QRCD. AraBERT is a model that is pre-trained on an MSA corpus, while CL-AraBERT is AraBERT with further pre-training on a Classical Arabic dataset. They found that the CL-AraBERT model outperformed AraBERT with 0.51 partial average precision (pAP)@10.

Other studies have used other techniques and language models to build Quran question-and-answer systems but have achieved worse results (Keleg and Magdy, 2022; Premasiri et al., 2022).

The studies in this approach applied the pre-trained language model for MRC task only while the questions-answer system consisting of two modules: Passage Retrieval (PR) and MRC. The first module will retrieve the Quran passages that likely contain the correct answer to the questions. The task of the second module is to extract the answer from the retrieved passages. In addition, the dataset size was very small, as the number of questions without duplication was 169.

2.2 GPT Research

Recently, many studies have investigated the performance of the GPT application in answering various types of questions. Katz et al. (2023) conducted an experiment to study the performance of GPT in passing the Uniform Bar Examination (UBE). Success on this exam is a condition for law practice in most states within the US. It consists of several components, including multiple-choice and essay items. GPT-4 (i.e. the fourth version of ChatGPT's underlying model) achieved good results with 297 points. Kung et al. (2023) suggested measuring the performance of GPT-3.5 in the United States Medical Licensing Exam (USMLE). The results showed that it was close to passing the test. Wood et al. (2023) recommend investigating the efficiency of GPT-3.5 using more than 28,000 questions from accounting tests. They found that the average results of the model were better than students when they considered partially correct answers. Kasai et al. (2023) proposed studying the performance of various versions of GPT using Japanese national medical licensing examinations for the last six years. The results showed that GPT-4 had the best performance and passed all exams. As far as we know, no study has assessed the GPT with Islamic questions.

3 Model

Based on the architecture of the pre-trained language models, they can be categorized into three categories: encoder-decoder (Raffel et al., 2020), decoder-only and encoder-only (Devlin et al., 2018). The GPT is considered a decoder-only model (Radford et al., 2018). The development of GPT has already passed through several generations: GPT-1 (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-3.5 (Ye et al., 2023) and GPT-4 (OpenAI, 2023). Various techniques have been used in these different versions, which have been reflected positively in the performance and development of the language model.

The version used for this study was GPT-4, the latest version of the GPT at the time of our research. This model is used directly and does not require any fine-tuning process. We used the Python API to retrieve answers from the model. Google Colab was used to run the experiment.

4 Datasets

In this section, we outline the datasets used in the study. This research focused on finding answers to questions from Islamic religious texts. In Islam, the Holy Quran is the primary source from which Muslims derive information to guide their lives.

4.1 Quran Question–Answer Dataset

The Holy Quran is God's word and Islam's basic book. Based on specific topics, the book is divided into 114 surahs. Each surah is a series of verses. The total number of verses in the Quran is 6,236 (Atwell et al., 2010).

As is well known in the field, there is a significant shortage of question–answer collections in low-resource languages, such as Arabic. There are only two available Quran question–answer datasets: the Annotated Corpus of Arabic Al-Quran Questions and Answers (AQQAC) and AyaTEC. The AQQAC consists of 1,224 question–answer pairs collected from the Islam – Quran & Tafseer website (Alqahtani, 2019).[1] A limitation of this dataset is that many answers are only from the words of scholars without any evidence from the Holy Quran. After excluding these types of answers, we identified only 611 question–answer pairs. A different approach was used to build AyaTEC. Its creators collected questions from the Internet, then several freelancers extracted the answers from the Quranic text and, finally, the questions and answers were reviewed by religious scholars (Malhas and Elsayed, 2020). AyaTEC consists of 169 questions, and each question may have more than one answer, so the number of records was 1,247. The nature of the combined AyaTEC and AQQAC questions was not considered challenging because they were direct and were a relatively small set. Therefore, we built the Quran question–answer pairs (QUQA) dataset.

4.2 QUQA Building Process

The QUQA creation methodology followed four steps: (1) corpus design, (2) source choice, (3) question and answer gathering and (4) data cleaning. First, the designs of AQQAC and AyaTEC were used as a basis when we planned the QUQA collection. It is available in a format using comma-separated values (CSVs) with UTF-8 encoding. Second, QUQA used the two available datasets (AQQAC and AyaTEC) and relevant published

---
[1] http://islamqt.com/

books. We chose books that fulfilled two requirements. Their answers had to include evidence from the Quran. Most books in this area include answers that involve only the expert's words without evidence. In addition, the owner of the book rights had to allow for our further publication of the data because we wanted to make the dataset available. The five books that met the requirements were 900 Questions and Answers in Managing the Verses of the Book (ALmuselli, 2020), 100 Quranic Questions and Answers (Alakeel, 2018), The Doctrine of Every Muslim in a Question and Answer (Zeno, 2004, 2007), Inference on Children's Treasure (Al-Wadi, 2016), and Prayer (1770) Question and Answer (Al Alami, 2022). Third, to collect the data, we wrote a programme that converts the sources from their original formats to the specific format used in our dataset. The AQQAC was available in one XLXS file, while AyaTEC consisted of three linked XML files. Some of the selected books were already available as files in a digital text format. Therefore, it was relatively easy to extract the questions and answers from them for our research. In contrast, the other books required an additional step to convert the scanned physical document into a digital text format. The fourth and final step involved cleaning the data by fixing incorrect information introduced during the process and removing duplicate data, noise and undesired items. In addition, the questions in this corpus were classified into three categories: fact, confirmation and descriptive. A fact was a simple answer to a question such as where, when, etc. A descriptive question requires a more complex and detailed answer. The confirmation category included yes/no questions. The data cleaning data process consisted of manual and automated tasks (the latter included using regular expressions to eliminate non-Arabic characters and removing extra spaces). This step enhanced the quality of the dataset.

The QUQA includes 2,189 questions, classified as 1778 single-answer and 411 multiple-answer questions. The answers in this dataset cover 47% of the Quran. The types of questions in the dataset are diverse and cover different topics, such as matters related to the afterlife, the foundations of faith, dealings between people and stories of the prophet. Examples of the QUQA dataset questions and their answers are shown in Table 1. This dataset is publicly available in our repository.[2]

---

[2] http://github.com/scsaln

Question ID	868
Question	مريم متفرغة للعبادة وطلب منها المزيد في الطاعات، أذكر الآية التي دلت على هذا المعنى؟. Mary is devoted to worship and is asked to do more acts of worship. Mention the verse that indicates this meaning?
Answer	يا مريم اقنتي لربك واسجدي واركعي مع الراكعين. Mary, be devoutly obedient to your Lord and prostrate and bow with those who bow [in prayer].

Table 1: Examples of the QUQA questions and their answers.

5 Evaluation

In response to the questions we supplied, the output generated by GPT-4 contains natural-sounding text and many series of Quranic verses, as shown in Table 2. In the output, the series may consist of one or more verses. Therefore, we evaluated the textual answer and the Quranic verses independently. The evaluation of the Quranic verses was both automatic and manual. First, we manually extracted the Quranic verses from the text because they were in different formats and put them in a standardised form. We wrote a programme to check whether the text of a Quranic verse was fake and then validated it as an answer to the question using the "golden" labelled dataset. We noticed that some of the verses mentioned in the GPT answer correctly addressed the question but were not mentioned in the golden dataset. Therefore, we checked all the answers and added unanticipated "found" correct answers to the dataset, as appropriate. For example, all the verses cited in the answer given in Table 2 are correct, but only the first verse was originally mentioned in the correct answer in the dataset. We evaluated the answers manually because the GPT answers may have a similar meaning to the golden answers but use different words. For example, in Question 2117 ("How long is full breastfeeding?") The answer is "سنتين"; but GPT answered with " حولين", these two words are synonymous in meaning two years.

We considered different retrieved series of Quranic verses in the answer as a ranked list and

Question ID	522
Question	ما عدد السموات ؟.
	How many heavens are there?
Answer	عدد السموات ذُكِرَ في القرآن الكريم بأنها سبع سماوات. يمكن العثور على هذا الرقم في عدة آيات مثل: ١) سورة البقرة (الآية ٢٩): آهُوَ الَّذِي خَلَقَ لَكُم مَّا فِي الْأَرْضِ جَمِيعًا ثُمَّ اسْتَوَىٰ إِلَى السَّمَاءِ فَسَوَّاهُنَّ سَبْعَ سَمَاوَاتٍ وَهُوَ بِكُلِّ شَيْءٍ عَلِيمٌ. ٢) سورة الإسراء (الآية ٤٤): آتُسَبِّحُ لَهُ السَّمَاوَاتُ السَّبْعُ وَالْأَرْضُ وَمَن فِيهِنَّ وَإِن مِّن شَيْءٍ إِلَّا يُسَبِّحُ بِحَمْدِهِ وَلَـٰكِن لَّا تَفْقَهُونَ تَسْبِيحَهُمْ إِنَّهُ كَانَ حَلِيمًا غَفُورًا
	The number of heavens mentioned in the Holy Quran is seven. This number can be found in several verses, such as (1) Surah Al-Baqarah (verse 29), "He it is Who created for you all that is on Earth. Then He Istawa (rose over) towards the heaven and made them seven heavens and He is the All-Knower of everything"; (2) Surah Al-Isra (verse 44), "The seven heavens and the Earth and whatever is in them exalt Him. And there is not a thing except that it exalts [Allah] by His praise, but you do not understand their [way of] exalting. Indeed, He is ever Forbearing and Forgiving."

Table 2: Example of GPT answers.

used F1@1, EM, F1 and pAP as evaluation metrics. F1@1 and EM are usually implemented on only the top predicted answer, which was the first series in our case. EM was a binary value. It was assigned a value of 1 when the first verse series was matched exactly with one of the gold verses or 0 if not. To compute F1, we measured the overlap between each series of verses and the golden answer and then took the average. If F1 was only computed for the first series, it was referred to as F1@1. The pAP measure was used to consider the rank of the correct answers in the retrieved list. If the system could retrieve the correct answers at the top of the list, then the score became higher. The above measures were computed for each question and then the average was taken. Malhas and El-sayed (2020, 2022) used these four measures to assess their Quran question–answer system. We dealt with the text as a bag of words and used EM and F1 to assess it. If the first sentence contained the exact answer, then the EM was assigned a value of 1. Otherwise, it was assigned a value of 0. The GPT-4 answers to all the questions are publicly available.[3]

---
[3] https://github.com/scsaln/GPT4

6 Results

In this section, we present the performance of GPT-4 with the QUQA dataset. We conducted three experiments. In each, we entered a particular type of question. The QUQA contains three kinds of question. The QUQA contains three kinds of questions: confirmation (103 questions), descriptive (1,621 questions) and fact (465 questions). The results summary of the Quranic series portion of the GPT-4 answers is shown in Table 3, while the produced-text portion results are shown in Table 4.

In general, GPT-4 did not achieve impressive results. We note that the first sentence and the first series of Quranic verses in its answers to the fact-type questions outperformed the other types with a 0.3 F1@1 score, 0.27 pAP and 0.25 EM score for the Quran verse portion and a 0.34 EM score for the regular text. For the entire regular text in the answer, GPT-4 achieved higher results in questions of the confirmation type, with a 0.29 F1 score. Additionally, GPT-4 obtained a high score with a 0.36 F1 score in the entire Quranic verses series portion of answering descriptive questions. The ranking of the Quranic verses series in the answers to All questions received a 0.23 pAP score, 0.32 F1, 0.19 EM, and 0.26 F1@1. While the text

Question Type	F1@1	EM	F1	pAP
All	0.26	0.19	0.32	0.23
Confirmation	0.22	0.13	0.29	0.18
Descriptive	0.26	0.19	**0.36**	0.24
Fact	**0.3**	**0.25**	0.32	**0.27**

Table 3: The evaluation results of the Quranic series portion of the GPT-4 answers.

Question Type	EM	F1
All	0.29	0.26
Confirmation	0.29	**0.29**
Descriptive	0.25	0.26
Fact	**0.34**	0.24

Table 4: The evaluation results of the produced-text portion of the GPT-4 answers.

portion received 0.26 F1 and 0.29 EM.

7 Analysis and Discussion

This section analyses and discusses the results obtained from the answers generated by GPT-4 in response to the QUQA questions.

Some incorrect answers appeared in the output, as illustrated in Table 5 and Table 6 (only the part of the answer containing the error is displayed due to space limitations). GPT-4 was unable to answer many questions, as shown in Example 1 an AI language model. Example 2 shows the generation of incomprehensible words. In Example 3, there is a contradiction, as it mentions a specific answer and then negates it in the same text. Example 4 shows an answer in which the text of the verse was correct, but the verse number and the name of the sura were incorrect. There were many answers in which the verse information was correct, but its text was incorrect. If GPT-4 found a matching word between the question and a verse, it retrieved the verse even when its meaning did not answer the question, as in Example 5. In Example 6, It failed to answer the question when the word of the correct answer verses and the word of the question were synonyms. It also faced difficulties if a word in the question and a verse were similar, but they were different in meaning due to context, as shown in Example 7. This led to incorrect processing of the question. For example, the word "شعب" in Arabic has the meaning of both "people" and "branches". The question in Example 7 means "branches of faith" based on the context of the sentence, but it has been processed by GPT-4 as equivalent to the "faithful people". In Example 8, the answer was composed of a series of verses. This seems to indicate that the model found it difficult to answer. In addition to the above examples, there were many answers in which Quranic verses were not mentioned, such as Question number 1. The other GPT-4 answers contain many related topics but do not answer the question directly, such as Question number 487. The answers to Question number 1 and Question number 487 are too long to present here but are available in our repository).

Ultimately, we wanted to research what some religious scholars believe is an unethical practice—in other words, using artificial intelligence to examine the Quran and found some specific issues with accuracy (Pournaras, 2023). Therefore, one must be careful when using language models and should not adopt any answer in this area without the consultation of an Islamic expert.

8 Conclusion

This paper presented the evaluation process of GPT in a Quranic question–answer task. We created a question–answer dataset called QUQA to benchmark this model. All questions in this dataset were sorted into three main categories: fact, confirmation and descriptive. The experiment revealed a generally weak GPT-4 performance. The performance of GPT-4 on the fact questions was better than the other types of questions with a 0.27 pAP score, 0.30 F1@1 score and 0.25 EM score. Our findings highlight the model limitations with regard to the Classical Arabic language and draw attention to the need for more research in this area. In the future, we will extend the corpus to include more challenging questions and use this dataset to build question–answering systems using other pre-trained models.

References

Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa M El-Makky, and Marwan Torki. 2014. Al-bayan: an arabic question answering system for the holy quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 57–64.

Asad Abdi, Shafaatunnur Hasan, Mohammad Arshi, Siti Mariyam Shamsuddin, and Norisma Idris. 2020. A question answering system in hadith using linguistic knowledge. *Computer Speech & Language*, 60:101023.

Example No.	1
Question ID	328
Question	Why did God Almighty choose Moses, peace be upon him, over other people during his time?
Part of the Answer	To view all the verses of the Holy Qur'an that relate to how God Almighty chose Moses, peace be upon him, over people at the time of his time, it is preferable to search for approved Qur'anic interpretations and explanations. Being an AI language model, I do not have the ability to accurately extract all the Quranic verses and can only provide general and epistemological answers based on the limited knowledge I have been trained on until 2021.
Example No.	2
Question ID	485
Question	What is the interpretation of Joseph, peace be upon him, of the king's dream?
Part of the Answer	The answer of Yusuf (peace be upon him) to the interpretation of the king's dream according to the verses (43-49) in Surat Yusuf: Yusuf said **ayn altftan sbaon ygrgsomagfnsAsdanh.**). . .
Example No.	3
Question ID	326
Question	Gold utensils are forbidden for use by a Muslim in this life, so is it the same in the hereafter?
Part of the Answer	**There is no explicit Qur'anic text that talks about vessels of gold in the hereafter**. However, it can be concluded that the pleasures and adornments that are forbidden in this life may be permissible in the hereafter because they are a reward for the believers for their good deeds. And in the Holy Qur'an there is a description of heaven and its bliss. In Surat Al-Zukhruf (43:71), **this verse indicates that the golden utensils will be used to serve the inhabitants of Paradise**, and they will be at the height of luxury....
Example No.	4
Question ID	382
Question	Why did Moses, peace be upon him, take the head of his brother Harun, peace be upon him, dragging him to Moses?
Part of the Answer	"...And when Moses returned to his people, angry and grieved, he said, "How wretched is that by which you have replaced me after [my departure]. Were you impatient over the matter of your Lord?" And he threw down the tablets and seized his brother by [the hair of] his head, pulling him toward him. [Aaron] said, "O son of my mother, indeed the people oppressed me and were about to kill me, so let not the enemies rejoice over me and do not place me among the wrongdoing people." **[Taha:94]**..

Table 5: Examples of wrong GPT answers, part 1.

Example No.	5
Question ID	385
Question	Why did the **desert-Arabs** not go out with the Messenger of God, may God's prayers and peace be upon him, to Mecca?
Part of the Answer	. . . But some of the **desert Arabs** believe in Allah and the Last Day, and look on their payments as pious gifts bringing them nearer to Allah and obtaining the prayers of the Messenger. Aye, indeed they bring them nearer (to Him): soon will Allah admit them to His Mercy: for Allah is Oft-forgiving, Most Merciful...[at-Tawbah:99]. . .
Golden Answer	Then who could prevent Allah at all if He intended for you harm or intended for you benefit? Rather, ever is Allah, with what you do, Acquainted [11]. But you thought that the Messenger and the believers would never return to their families, ever, and that was made pleasing in your hearts. And you assumed an assumption of evil and became a people ruined."[12].[Sura Al-Fat'h].

Example No.	6
Question ID	534
Question	What is the sign of the **torment coming** to the people of Noah, peace be upon him?
Part of the Answer	. . . And it was said, "O Earth, swallow your water, and O sky, withhold [your rain]." And the water subsided, and the matter was accomplished, and the ship came to rest on the [mountain of] Judiyy. And it was said, "Away with the wrongdoing people."[Hud:44]. . .
Golden Answer	[So it was], until when **Our command came** and the oven overflowed, We said, "Load upon the ship of each [creature] two mates and your family, except those about whom the word has preceded, and [include] whoever has believed." But none had believed with him, except a few.

Example No.	7
Question ID	1892
Question	Describe some **branches of faith**?
Part of the Answer	In the Holy Qur'an, **many peoples of faith**are referred to, including: 1. The Children of Israel
Golden Answer	... **Branches of faith** are Belief in God, His angels, His books, His messengers, and the Last Day...

Example No.	8
Question ID	291
Question	Describe the **percentage** of people entering heaven from the **forerunners** rank?
Part of the Answer	There is **no specific number** for the percentage of people entering Paradise from the rank of the **"forerunners "**in the Holy Quran....
Golden Answer	And **the forerunners, the forerunners** [10]. These are the nearest [11]. In the Gardens of Bliss [12]. **A group from the ancients** [13]. And **a few from the latecomers** [14].[Surah Al-Waqia]

Table 6: Examples of wrong GPT answers, part 2.

Basem Ahmed, Motaz Saad, and Eshrag A Refaee. 2022. Qqateam at qur'an qa 2022: Fine-tunning arabic qa models for qur'an qa task. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 130–135.

Faisal bin Misfer bin Moawad Al Alami. 2022. *Prayer (1770) Question and Answer*.

Faisal bin Misfer bin Moawad Al-Wadi. 2016. *Inference on children's treasure*. Dar Knoz Al-Islam.

Fouzia Alakeel. 2018. *Quranic questions and answer*.

Duraid ALmuselli. 2020. *900 questions and answers in managing the verses of the book*. Altafseer, Erbil.

Mohammad Mushabbab A Alqahtani. 2019. *Quranic Arabic semantic search model based on ontology of concepts*. Ph.D. thesis, University of Leeds.

Abdullah Alsaleh, Saud Althabiti, Ibtisam Alshammari, Sarah Alnefaie, Sanaa Alowaidi, Alaa Alsaqer, Eric Atwell, Abdulrahman Altahhan, and Mohammad Ammar Alsalka. 2022. Lk2022 at qur'an qa 2022: Simple transformers model for finding answers to questions from qur'an. In *Proceedings of the OSACT 2022 Workshop*, pages 120–125. ELRA European Language Resources Association.

Shatha Altammami and Eric Atwell. 2022. Challenging the transformer-based models with a classical arabic dataset: Quran and hadith. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1462–1471.

Eric Atwell, Nizar Habash, Bill Louw, Bayan Abu Shawar, Tony McEnery, Wajdi Zaghouani, and Mahmoud El-Haj. 2010. Understanding the quran: A new grand challenge for computer science and artificial intelligence. *ACM-BCS Visions of Computer Science 2010*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mohammed ElKomy and Amany M Sarhan. 2022. Tce at qur'an qa 2022: Arabic language question answering over holy qur'an using a post-processed ensemble of bert-based models. *arXiv preprint arXiv:2206.01550*.

Bothaina Hamoud and Eric Atwell. 2016. Quran question and answer corpus for data mining with weka. In *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, pages 211–216. IEEE.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Amr Keleg and Walid Magdy. 2022. Smash at qur'an qa 2022: Creating better faithful data splits for low-resourced question answering scenarios. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 136–145.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–21.

Rana Malhas and Tamer Elsayed. 2022. Arabic machine reading comprehension on the holy qur'an using cl-arabert. *Information Processing & Management*, 59(6):103068.

Hajer Maraoui, Kais Haddar, and Laurent Romary. 2021. Arabic factoid question-answering system for islamic sciences using normalized corpora. *Procedia Computer Science*, 192:69–79.

Ali Mostafa and Omar Mohamed. 2022. Gof at qur'an qa 2022: Towards an efficient question answering for the holy qu'ran in the arabic language using deep learning-based approach. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 104–111.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Evangelos Pournaras. 2023. Science in the era of chatgpt, large language models and ai: Challenges for research ethics review and how to respond. *arXiv preprint arXiv:2305.15299*.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. Dtw at qur'an qa 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. *arXiv preprint arXiv:2205.06025*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

David A Wood, Muskan P Achhpilia, Mollie T Adams, Sanaz Aghazadeh, Kazeem Akinyele, Mfon Akpan, Kristian D Allee, Abigail M Allen, Elizabeth D Almer, Daniel Ames, et al. 2023. The chatgpt artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education*, pages 1–28.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Muhammad bin Jamil Zeno. 2004. *The abbreviation of the Islamic belief from the Qur'an and Sunnah*.

Muhammad bin Jamil Zeno. 2007. *The doctrine of every Muslim in a question and answer*.