QATAR UNIVERSITY

COLLEGE OF ENGINEERING

ARABIC QUESTION ANSWERING ON THE HOLY QUR’AN

BY

RANA R. MALHAS

A Dissertation Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Doctorate of Philosophy in Computer Science

January 2023

COMMITTEE PAGE

The members of the Committee approve the Dissertation of

Rana R. Malhas defended on 27/12/2022.

Dr. Tamer Elsayed
Dissertation Supervisor

Prof. Eric Atwell (University of Leeds)
Committee Member

Prof. Abdelaziz Bouras (Qatar University)
Committee Member

Prof. Junaid Qadir (Qatar University)
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

Malhas, Rana, R., Doctorate : January : 2023, Doctorate of Philosophy in Computer Science

Title: Arabic Question Answering on the Holy Qur'an

Supervisor of Dissertation: Dr. Tamer Elsayed.

In this dissertation, we address the need for an intelligent *machine reading at scale* (MRS) Question Answering (QA) system on the Holy Qur'an, given the permanent interest of inquisitors and knowledge seekers in this sacred and fertile knowledge resource. We adopt a pipelined *Retriever-Reader* architecture for our system to constitute (to the best of our knowledge) the first extractive MRS QA system on the Holy Qur'an. We also construct *QRCD* as the first extractive Qur'anic Reading Comprehension Dataset, composed of 1,337 question-passage-answer triplets for 1,093 question-passage pairs that comprise single-answer and multi-answer questions in modern standard Arabic (MSA). We then develop a sparse bag-of-words passage retriever over an index of Qur'anic passages expanded with Qur'an-related MSA resources to help in bridging the gap between questions posed in MSA and their answers in Qur'anic Classical Arabic (CA). Next, we introduce CLassical AraBERT (CL-AraBERT for short), a new AraBERT-based pre-trained model that is further pre-trained on about 1.05B-word Classical Arabic dataset (after being initially pre-trained on MSA datasets), to make it a better fit for NLP tasks on CA text such as the Holy Qur'an. We leverage cross-lingual transfer learning from MSA to CA, and fine-tune CL-AraBERT as a reader using a couple of MSA-based MRC datasets followed by fine-tuning it on our *QRCD* dataset, to bridge the above MSA-to-CA gap, and circumvent the lack of MRC datasets in CA. Finally, we integrate the retriever and reader components of the end-to-end QA system such that the top k retrieved answer-bearing passages to a given question are fed to the fine-tuned CL-AraBERT reader for answer extraction. We first evaluate the retriever and the reader components independently, before evaluating the end-to-end QA system using *Partial Average Precision* (pAP). We introduce pAP as an adapted version of the traditional rank-based Average Precision measure, which integrates partial matching in the evaluation over multi-answer and single-answer questions.

Our experiments show that a passage retriever over a BM25 index of Qur'anic passages expanded with two MSA resources significantly outperformed a baseline retriever over an index of Qur'anic passages only. Moreover, we empirically show that the fine-tuned CL-AraBERT reader model significantly outperformed the similarly fine-tuned AraBERT model, which is the baseline. In general, the CL-AraBERT reader performed better on single-answer questions in comparison to multi-answer questions. Moreover, it has also outperformed the baseline over both types of questions. Furthermore, despite the integral contribution of fine-tuning with the MSA datasets in enhancing the performance of the readers, relying exclusively on those datasets (without MRC datasets in CA, e.g., *QRCD*) may not be sufficient for our reader models. This finding demonstrates the relatively high impact of the *QRCD* dataset (despite its modest size). As for the QA system, it consistently performed better on single-answer questions in comparison to multi-answer questions. However, our experiments provide enough evidence to suggest that a native BERT-based model architecture fine-tuned on the MRC task may not be intrinsically optimal for multi-answer questions.

DEDICATION

To my family for their support and patience.

ACKNOWLEDGMENTS

I would like to warmly thank my advisor, Dr. Tamer Elsayed for his valuable and continuous support and dedication. His insightful scientific vision and resourceful guidance helped me through the challenges of my PhD journey.

I would like to thank Dr. Ahmad Shukri, Professor of Tafseer and Qur'anic Sciences at Qatar University for his scholarly advice throughout the annotation process of the answers extracted from the Holy Qur'an.

I appreciatively acknowledge the support of Qatar University for partially funding the research work in this dissertation through the two grants numbered: QUST-1-CENG-2018-24 and QUST-2-CENG-2020-20.

I would also like to thank the TensorFlow Research Cloud (TFRC) team for their valuable support in providing free access to their cloud TPUs for training the models in this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The Qur'an is sacredly held by more than 1.9 billion Muslims across the world.[1] It is the major source of knowledge, teachings, wisdom, and legislation in Islam, in addition to its inclusion of scientific [117] knowledge that is beneficial to mankind. All this encompassed knowledge makes the Holy Qur'an a rich and fertile source for Muslim and non-Muslim knowledge-seekers pursuing answers to questions raised for learning, out of curiosity, or skepticism. The Holy Qur'an is composed of 114 chapters (Suras) and 6236 verses (Ayas) of different lengths, with a total of about 80k Arabic words. The words, revealed more than 1,400 years ago, are in *Classical Arabic* (CA) [28]. It is a phenomenal yet challenging document collection due to its long-chained anaphoric-structures across the verses of the same chapter, in addition to the large diversity of its topic categories that are scattered in different positions of the Qur'an. Moreover, a qur'anic verse may relate to one or more topics, and the same topic may be tackled in different chapters/verses, but in variant contexts [85]. We denote this challenging feature in the Holy Qur'an by "unstructured topic diversity".

Understanding the Holy Qur'an [28] and its encompassed knowledge is essential to Muslims and the societies and communities they thrive in, not only because it touches every aspect of their lives, but also to clear any misconceptions towards Islam that may arise among members of their Muslim or non-Muslim communities. As such, there will always be a need for intelligent question answering (QA) systems on the Holy Qur'an that can address the information needs of its curious as well as skeptical users (knowledge-seekers). In fact, a recent systematic review on Arabic NLP for Qur'anic research [35] has explicitly identified the need for intelligent systems to answer the questions of Muslims and non-Muslims as a required open issue for future research

[1] `https://en.wikipedia.org/wiki/Islam_by_country`

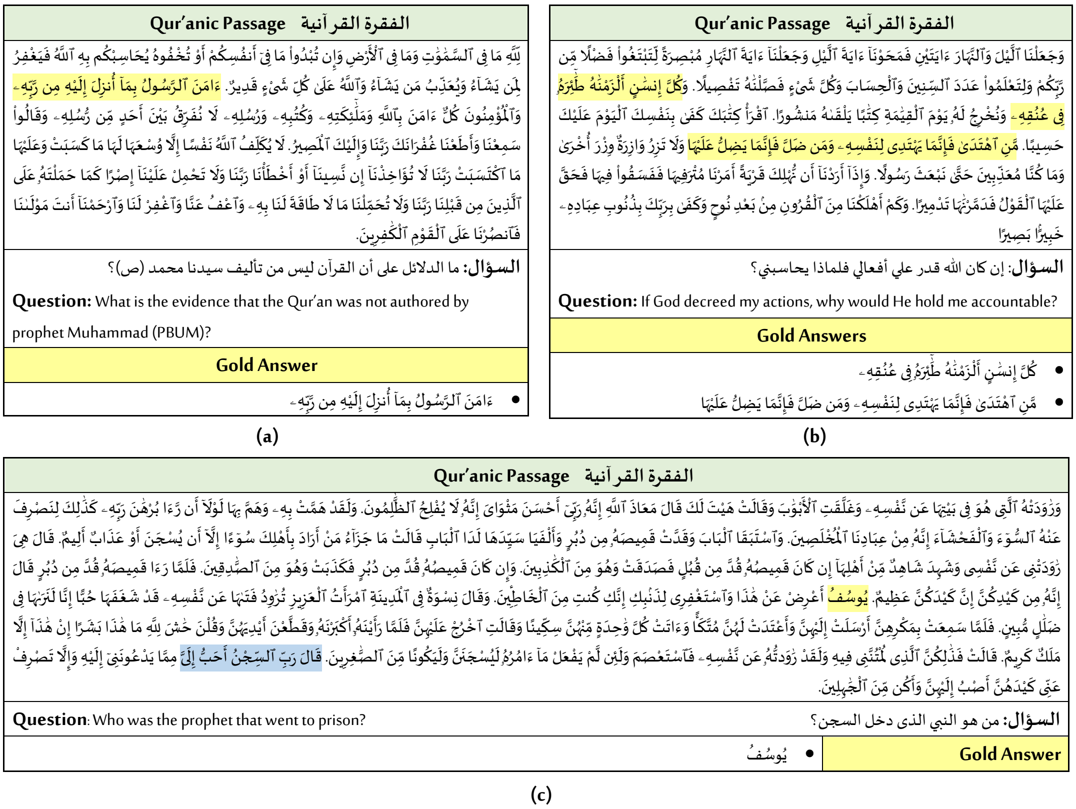
directions. Moreover, the study revealed that most of the prevalent semantic-based search systems for Qur'an are concept/topic-oriented rather than user-oriented, because they "are designed from a topic perspective and not from a user perspective" [35].

In general, QA systems are broadly classified in the literature as either *knowledge base* (KB)-QA or *textual*-QA, depending on the information sources they use in extracting or generating the answers [29], [148]. KB-QA typically mines the answers from manually constructed KB structures, while textual-QA extracts or generates them from unstructured text documents (including those on the world wide web). Textual QA approaches are recently being formulated as machine reading comprehension (MRC) tasks [20], [148]. In the 1970s, MRC was initially perceived as the ideal apparatus to evaluate the task of language understanding by computer systems [38]. Given a passage of text, a machine reading comprehension system should read this passage and answer comprehension questions about it [38]. After being dormant for decades, the MRC field witnessed a resurgence that was mainly attributed to the development of large reading comprehension datasets [67], [75], [111], which enabled the training of deep learning neural MRC systems. These datasets are readily suitable for MRC tasks because each question-answer pair is coupled with the passage(s) or document to which the answer was extracted/generated from. As such, they include tuples of question-passage-answer triplets [38]. Moreover, the advent and phenomenal success of transformer-based pre-trained language models, e.g., BERT [48], RoBERTa [83] and XLNet [141], have further escalated the rate at which the field of neural MRC was progressing.

Interestingly, the perception towards the MRC task has evolved from being a question answering (QA) task over a closed piece of text into an integral component of modern AI systems, such as *machine reading at scale* (also called Open-domain QA) that adopts the "*Retriever-Reader*" architecture [38], [39], [99], [140], [148]. This is not to demote the importance of reading comprehension in closed settings (over a given text), where the systems are relieved from the task of passage retrieval to purely focus on inference and reasoning for answer extraction [104] or answer generation [34], [74], [148]. In general, machine reading at scale (MRS) or Open-domain QA (OpenQA for short) are used interchangeably in the literature to refer to answering a given question without specifying the context to which the answers will be extracted/generated from (as opposed to the traditional MRC task where the context to which the answer is extracted or generated from, is specified). Thus, MRS (or OpenQA) enjoys a wider scope of application over the world wide web or a local document repository [148], which makes it suitable for application over a closed-domain like the Holy Qur'an.

In a nutshell, the problem we address in this dissertation is: given a question in *modern standard Arabic* (MSA), a QA system should return a ranked list of answers (as extracted spans of text) from the Holy Qur'an. Moreover, the QA system should aim to find the answer to the given MSA question anywhere in the Holy Qur'an. The question can be factoid or non-factoid. Factoid questions mainly include "who", "when", "where" and "how long/many" questions, while non-factoid questions mainly include "why", "describe", and "evidence" questions.[2] Figure 1.1 exhibits examples of questions that reflect some of the challenges of this problem. For example, it is of paramount importance for such a QA system to address the challenge of bridging the gap between the questions being in MSA and the answers being in Qur'anic Classical Arabic; we denote this gap by the *MSA-to-CA gap* for short (Figure 1.1-(b) presents an example of this gap). In general, due to the literary style of Qur'anic text, answers to non-factoid

[2] Evidence questions mainly include "what is the ruling", "what indications/evidence" and "yes/no" questions. For example, answer(s) to a "yes/no" question is drawn from verses that provide evidence that asserts or negates that question.



Qur'anic Passage الفقرة القرآنية

لِّلَّهِ مَا فِى ٱلسَّمَٰوَٰتِ وَمَا فِى ٱلْأَرْضِ وَإِن تُبْدُوا۟ مَا فِىٓ أَنفُسِكُمْ أَوْ تُخْفُوهُ يُحَاسِبْكُم بِهِ ٱللَّهُ فَيَغْفِرُ لِمَن يَشَآءُ وَيُعَذِّبُ مَن يَشَآءُ وَٱللَّهُ عَلَىٰ كُلِّ شَىْءٍ قَدِيرٌ. ءَامَنَ ٱلرَّسُولُ بِمَآ أُنزِلَ إِلَيْهِ مِن رَّبِّهِۦ وَٱلْمُؤْمِنُونَ كُلٌّ ءَامَنَ بِٱللَّهِ وَمَلَٰٓئِكَتِهِۦ وَكُتُبِهِۦ وَرُسُلِهِۦ لَا نُفَرِّقُ بَيْنَ أَحَدٍ مِّن رُّسُلِهِۦ وَقَالُوا۟ سَمِعْنَا وَأَطَعْنَا غُفْرَانَكَ رَبَّنَا وَإِلَيْكَ ٱلْمَصِيرُ. لَا يُكَلِّفُ ٱللَّهُ نَفْسًا إِلَّا وُسْعَهَا لَهَا مَا كَسَبَتْ وَعَلَيْهَا مَا ٱكْتَسَبَتْ رَبَّنَا لَا تُؤَاخِذْنَآ إِن نَّسِينَآ أَوْ أَخْطَأْنَا رَبَّنَا وَلَا تَحْمِلْ عَلَيْنَآ إِصْرًا كَمَا حَمَلْتَهُۥ عَلَى ٱلَّذِينَ مِن قَبْلِنَا رَبَّنَا وَلَا تُحَمِّلْنَا مَا لَا طَاقَةَ لَنَا بِهِۦ وَٱعْفُ عَنَّا وَٱغْفِرْ لَنَا وَٱرْحَمْنَآ أَنتَ مَوْلَىٰنَا فَٱنصُرْنَا عَلَى ٱلْقَوْمِ ٱلْكَٰفِرِينَ.

السؤال: ما الدلائل على أن القرآن ليس من تأليف سيدنا محمد (ص)؟

Question: What is the evidence that the Qur'an was not authored by prophet Muhammad (PBUM)?

Gold Answer

- ءَامَنَ ٱلرَّسُولُ بِمَآ أُنزِلَ إِلَيْهِ مِن رَّبِّهِۦ

(a)

Qur'anic Passage الفقرة القرآنية

وَجَعَلْنَا ٱلَّيْلَ وَٱلنَّهَارَ ءَايَتَيْنِ فَمَحَوْنَآ ءَايَةَ ٱلَّيْلِ وَجَعَلْنَآ ءَايَةَ ٱلنَّهَارِ مُبْصِرَةً لِّتَبْتَغُوا۟ فَضْلًا مِّن رَّبِّكُمْ وَلِتَعْلَمُوا۟ عَدَدَ ٱلسِّنِينَ وَٱلْحِسَابَ وَكُلَّ شَىْءٍ فَصَّلْنَٰهُ تَفْصِيلًا. وَكُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ وَنُخْرِجُ لَهُۥ يَوْمَ ٱلْقِيَٰمَةِ كِتَٰبًا يَلْقَىٰهُ مَنشُورًا. ٱقْرَأْ كِتَٰبَكَ كَفَىٰ بِنَفْسِكَ ٱلْيَوْمَ عَلَيْكَ حَسِيبًا. مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا وَلَا تَزِرُ وَازِرَةٌ وِزْرَ أُخْرَىٰ وَمَا كُنَّا مُعَذِّبِينَ حَتَّىٰ نَبْعَثَ رَسُولًا. وَإِذَآ أَرَدْنَآ أَن نُّهْلِكَ قَرْيَةً أَمَرْنَا مُتْرَفِيهَا فَفَسَقُوا۟ فِيهَا فَحَقَّ عَلَيْهَا ٱلْقَوْلُ فَدَمَّرْنَٰهَا تَدْمِيرًا. وَكَمْ أَهْلَكْنَا مِنَ ٱلْقُرُونِ مِنۢ بَعْدِ نُوحٍ وَكَفَىٰ بِرَبِّكَ بِذُنُوبِ عِبَادِهِۦ خَبِيرًۢا بَصِيرًا

السؤال: إن كان الله قدر علي أفعالي فلماذا يحاسبني؟

Question: If God decreed my actions, why would He hold me accountable?

Gold Answers

- كُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ
- مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا

(b)

Qur'anic Passage الفقرة القرآنية

وَرَٰوَدَتْهُ ٱلَّتِى هُوَ فِى بَيْتِهَا عَن نَّفْسِهِۦ وَغَلَّقَتِ ٱلْأَبْوَٰبَ وَقَالَتْ هَيْتَ لَكَ قَالَ مَعَاذَ ٱللَّهِ إِنَّهُۥ رَبِّىٓ أَحْسَنَ مَثْوَاىَ إِنَّهُۥ لَا يُفْلِحُ ٱلظَّٰلِمُونَ. وَلَقَدْ هَمَّتْ بِهِۦ وَهَمَّ بِهَا لَوْلَآ أَن رَّءَا بُرْهَٰنَ رَبِّهِۦ كَذَٰلِكَ لِنَصْرِفَ عَنْهُ ٱلسُّوٓءَ وَٱلْفَحْشَآءَ إِنَّهُۥ مِنْ عِبَادِنَا ٱلْمُخْلَصِينَ. وَٱسْتَبَقَا ٱلْبَابَ وَقَدَّتْ قَمِيصَهُۥ مِن دُبُرٍ وَأَلْفَيَا سَيِّدَهَا لَدَا ٱلْبَابِ قَالَتْ مَا جَزَآءُ مَنْ أَرَادَ بِأَهْلِكَ سُوٓءًا إِلَّآ أَن يُسْجَنَ أَوْ عَذَابٌ أَلِيمٌ. قَالَ هِىَ رَٰوَدَتْنِى عَن نَّفْسِى وَشَهِدَ شَاهِدٌ مِّنْ أَهْلِهَآ إِن كَانَ قَمِيصُهُۥ قُدَّ مِن قُبُلٍ فَصَدَقَتْ وَهُوَ مِنَ ٱلْكَٰذِبِينَ. وَإِن كَانَ قَمِيصُهُۥ قُدَّ مِن دُبُرٍ فَكَذَبَتْ وَهُوَ مِنَ ٱلصَّٰدِقِينَ. فَلَمَّا رَءَا قَمِيصَهُۥ قُدَّ مِن دُبُرٍ قَالَ إِنَّهُۥ مِن كَيْدِكُنَّ إِنَّ كَيْدَكُنَّ عَظِيمٌ. يُوسُفُ أَعْرِضْ عَنْ هَٰذَا وَٱسْتَغْفِرِى لِذَنۢبِكِ إِنَّكِ كُنتِ مِنَ ٱلْخَاطِـِٔينَ. وَقَالَ نِسْوَةٌ فِى ٱلْمَدِينَةِ ٱمْرَأَتُ ٱلْعَزِيزِ تُرَٰوِدُ فَتَىٰهَا عَن نَّفْسِهِۦ قَدْ شَغَفَهَا حُبًّا إِنَّا لَنَرَىٰهَا فِى ضَلَٰلٍ مُّبِينٍ. فَلَمَّا سَمِعَتْ بِمَكْرِهِنَّ أَرْسَلَتْ إِلَيْهِنَّ وَأَعْتَدَتْ لَهُنَّ مُتَّكَـًٔا وَءَاتَتْ كُلَّ وَٰحِدَةٍ مِّنْهُنَّ سِكِّينًا وَقَالَتِ ٱخْرُجْ عَلَيْهِنَّ فَلَمَّا رَأَيْنَهُۥٓ أَكْبَرْنَهُۥ وَقَطَّعْنَ أَيْدِيَهُنَّ وَقُلْنَ حَٰشَ لِلَّهِ مَا هَٰذَا بَشَرًا إِنْ هَٰذَآ إِلَّا مَلَكٌ كَرِيمٌ. قَالَتْ فَذَٰلِكُنَّ ٱلَّذِى لُمْتُنَّنِى فِيهِ وَلَقَدْ رَٰوَدتُّهُۥ عَن نَّفْسِهِۦ فَٱسْتَعْصَمَ وَلَئِن لَّمْ يَفْعَلْ مَآ ءَامُرُهُۥ لَيُسْجَنَنَّ وَلَيَكُونًا مِّنَ ٱلصَّٰغِرِينَ. قَالَ رَبِّ ٱلسِّجْنُ أَحَبُّ إِلَىَّ مِمَّا يَدْعُونَنِىٓ إِلَيْهِ وَإِلَّا تَصْرِفْ عَنِّى كَيْدَهُنَّ أَصْبُ إِلَيْهِنَّ وَأَكُن مِّنَ ٱلْجَٰهِلِينَ.

Question: Who was the prophet that went to prison?

السؤال: من هو النبي الذى دخل السجن؟

Gold Answer

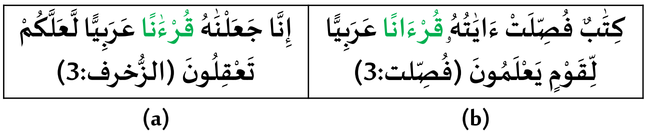
- يُوسُفُ

(c)

Figure 1.1. Example MRC questions and answers. (a) A non-factoid question with an evidence-based answer that is a single span of text. (b) A non-factoid question with two evidence-based answers (spans). It also showcases the MSA-to-CA gap, where the first answer includes the word "ta'erahu" which means "his bird" in MSA, while in Qur'anic CA, it means "his deeds and their implications on his happiness or misery". (c) A factoid question whose answer showcases a relatively long anaphoric-structure. Text highlighted in blue is the reference expression to the preceding antecedent (answer) highlighted in yellow.

questions are mostly evidence-based (Figure 1.1-(a) and (b)), while answers to factoid questions are likely to require some form of coreference resolution (Figure 1.1-(c)). Consequently, our QA task on the Qur'an requires multi-verse reasoning. Moreover, the evidence-based nature of the answers supports our rationale for formulating the problem as a rank-based task, because it would tend to better address the information needs of inquisitors, who would rather see *all* answers. Hence, the QA system should address the challenges posed by the Arabic language (in its two forms), in addition to those posed by the Qur'anic text, which include long-chained anaphoric structures, and unstructured topic diversity (as mentioned above).

Although CA and MSA share the same morphology and syntax characteristics, they mainly differ in lexis, where contemporary western words found their way into MSA through translation or transliteration and obsolete words were dropped [98]. Nevertheless, CA remains richer in lexis [121], which widens the MSA-to-CA gap. This gap is further compounded due to the rather sporadic non-conformity of the Holy Qur'an's Uthmani orthography[3] to Classical Arabic (as shown in Figure 1.2), which is an open issue in Qur'anic NLP research [35].



(a)	(b)
إِنَّا جَعَلْنَٰهُ قُرْءَٰنًا عَرَبِيًّا لَّعَلَّكُمْ تَعْقِلُونَ (الزُّخرف:3)	كِتَٰبٌ فُصِّلَتْ ءَايَٰتُهُۥ قُرْءَانًا عَرَبِيًّا لِّقَوْمٍ يَعْلَمُونَ (فُصِّلت:3)

Figure 1.2. Examples of the non-conformity of the Qur'an orthography to Classical Arabic. In (a) and (b), we exhibit two verses showing words whose "dagger alif" (or "alif khanjariyah") replace the traditional long vowel "alif". In some cases, the same word (e.g. the word "Qur'an" in green) may appear in different verses using either one of the "alif" forms.

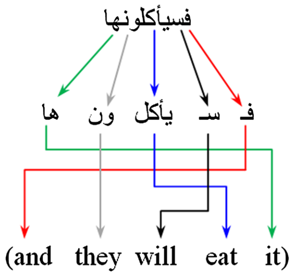


Figure 1.3. An example of high inflection in one single Arabic word [4].

In general, the Arabic language in its two forms (MSA and CA) poses challenges to any Natural Language Processing (NLP) task (including QA and MRC). It is a highly inflectional language, which makes it extensively morphological; for example, a single word may have several morphemes as shown in Figure 1.3. Other challenges include

[3] Al-rasm al-Uthmani (or rasm al-mushaf) is the convention adopted for writing the Qur'anic text during the ruling of Caliph Uthman bin Affan [30], [35].

the absence of capital letters and lack of diacritics in MSA. Diacritics are important because they disambiguate the meaning hence understanding of Arabic text, given that a change in the diacritics of a single letter in a word may utterly change the meaning of that word. Figure 1.4 presents an example of an Arabic word that could mean “science” or “flag” depending on what diacritics were used to annotate its letters. Although the Holy Qur’an is heavily diacritized, most NLP tasks over digital Qur’anic text resort to normalization by removing diacritics in the preprocessing stage [35]. In such cases, only the context of words is used to disambiguate their intended meaning, which poses additional challenges to NLP systems.

We do acknowledge the sensitivity and importance of maintaining the Uthmani orthography style of the Holy Qur’an [29] in printed and digital form. This may be possible and even important to some NLP tasks (such as part-of-speech tagging, segmentation among others), but it is a big hindrance to search, QA and MRC tasks over digital content, where normalization of text is the status quo for achieving better performance. Another aspect of equal importance, is the need for involving Qur’an scholars/experts to make sure that inputs/outputs of any NLP task on the Qur’an are not astray from the consensus of early scholars [35].

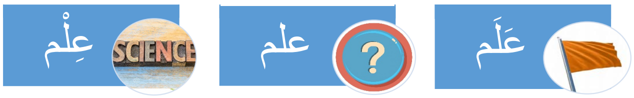


Figure 1.4. An example of how diacritics can change the meaning of an Arabic word.

The final challenge to tackle is the scarcity of Arabic QA resources for training and evaluation (in comparison to English QA resources, for example). The majority of prevalent resources are in modern standard Arabic, while classical Arabic QA resources received little attention. Furthermore, the absence of *fully-reusable test collections* for

Arabic QA and MRC tasks on the Holy Qur'an has impeded the possibility of fairly comparing the performance of systems in that domain. In general, a *test collection* is typically composed of a document collection[4] (the Holy Qur'an in our case), a set of queries (questions), and their relevance judgments [80], [132] (i.e., the gold answers or the passages that comprise them, in our case). For a QA test collection to be *reusable*, it must incorporate a non-trivial coverage of relevant answers to the respective questions [80]. Optimally, when building a QA test collection for the Holy Qur'an (or any religious book for that matter), it should be *fully-reusable* by aiming to include *all* relevant answers to each question.

In the following sections of this chapter, we formally introduce the problem statement, and an overview of the approach adopted in this dissertation. Then we present the six research questions that this work was designed to address, before concluding this chapter with the main contributions of this research work.

1.1. Problem Statement

Given a (factoid or non-factoid) question in MSA, a Question Answering system should return a ranked list of answers (spans) from the Holy Qur'an to the given question. We address the complexity of the problem by partitioning it into two sub-problem statements.

1. Given a question in MSA, a retrieval/search system should retrieve the top k answer-bearing passages from the Holy Qur'an.

2. Given a question-passage pair, a Machine Reading Comprehension system should extract the best answer(s) to the given MSA question from the accompa-

[4] In information retrieval, researches use the term "document collection" or "collection" to refer to a corpus or dataset [82]; we use these terms interchangeably.

nying Qur'anic passage. If the question has more than one answer in the passage, the system is expected to extract *all* answers.

1.2. Approach Overview

In this work, we address the need for an intelligent *machine reading at scale* QA system over the Holy Qur'an, given the permanent interest of inquisitors and knowledge seekers in this fertile knowledge resource. Inspired by the recent surge in the literature to adopt the *Retriever-Reader* architecture for machine reading at scale (MRS) [38], [39], [99], [148], where the *Retriever* is typically an information retrieval system, and the *Reader* is typically a neural MRC system, we adopt the same architecture for developing (to the best of knowledge) the first MRS QA system on the Holy Qur'an. The QA system is expected to receive a question in MSA and aims to find the answer anywhere in the Holy Qur'an. With the success of transformer-based pre-trained language models [82], [148], we were eager to develop an Arabic BERT-based reader over the Qur'an. Not demoting the importance of the retriever component, we have also developed a sparse bag-of-words passage retriever with document expansion using MSA resources, as it is of paramount importance for both components to address the MSA-to-CA gap.

To address the absence of *fully-reusable* test collections for Arabic QA on the Holy Qur'an, we introduce *AyaTEC* [85], a verse-based and fully-reusable test collection for evaluating Arabic question answering systems on the Holy Qur'an. It can also serve as a training CA resource.[5] *AyaTEC* includes 207 questions (with their corresponding 1,762 answers) covering 11 topic categories of the Holy Qur'an that target the information needs of both curious and skeptical users. To the best of effort, the answers to the

[5] With "*ayah*" being a "qur'anic verse" in Arabic, it inspired the naming of our test collection.

questions (each represented as a sequence of verses) in *AyaTEC* are exhaustive; i.e., all the qur'anic verses that directly answer the questions were exhaustively extracted and annotated.

To support the training of the reader component of our QA system, we extend *AyaTEC* to develop *QRCD* as the first extractive Qur'anic Reading Comprehension Dataset that adopts the same format of SQuAD v1.1 [111]. Each of the two datasets serves as a common experimental test-bed to fairly compare systems, as well as a Qur'anic training resource for QA and MRC models. *Extractive* MRC refers to the task of span prediction, where the answer is a specific span of text extracted (rather than generated) from passages accompanying a question [34], [38], [148]. *QRCD* is composed of 1,337 question-passage-answer triplets for 1,093 question-passage pairs. The MSA questions in *QRCD* are of two types, single-answer and multi-answer questions; each question is coupled with its corresponding curated passage(s) from the Qur'an. Answers to multi-answer questions are composed of two or more components. Thus, *QRCD* presents an additional challenge to QA and MRC tasks.

In Figure 1.5, we exhibit an overview of the pipelined retriever-reader architecture of the QA system. It is developed such that it attempts to address the challenges of the Qur'anic text, the Arabic language, and the nature of the QA task that were presented at the beginning of this chapter. Given a question in MSA, the retriever component searches an inverted index of Qur'anic passages that are expanded with two MSA resources, to help in bridging the gap between the questions being in MSA and the answers being in Qur'anic Classical Arabic. The first resource is Al-Tafseer Al-Muyassar [1], which is a simple interpretation of the Holy Qur'an in MSA, while the second is a Dictionary of Qur'anic words with their meaning in MSA [84]. The top K scoring passages that are returned by the Okapi BM25 [113] index search are then passed to the Arabic BERT-based reader as Qur'anic-only passages. The reader in turn extracts and returns the best answers from *all* these passages ranked by their normalized scores.

The reader was developed by first further pre-training AraBERT [23] using about 1.05B-word Classical Arabic corpus to complement the MSA resources used in pre-training the initial model, and make it a better fit for our task. We denote this model by CL-AraBERT (CLassical AraBERT for short). Finally, we fine-tuned CL-AraBERT as a reader using two MRC datasets in MSA, prior to fine-tuning it using our *QRCD* dataset. We cast the problem as a cross-lingual transfer learning task from MSA to CA not only to bridge the MSA-to-CA gap but also to overcome the modest size of the *QRCD* dataset.

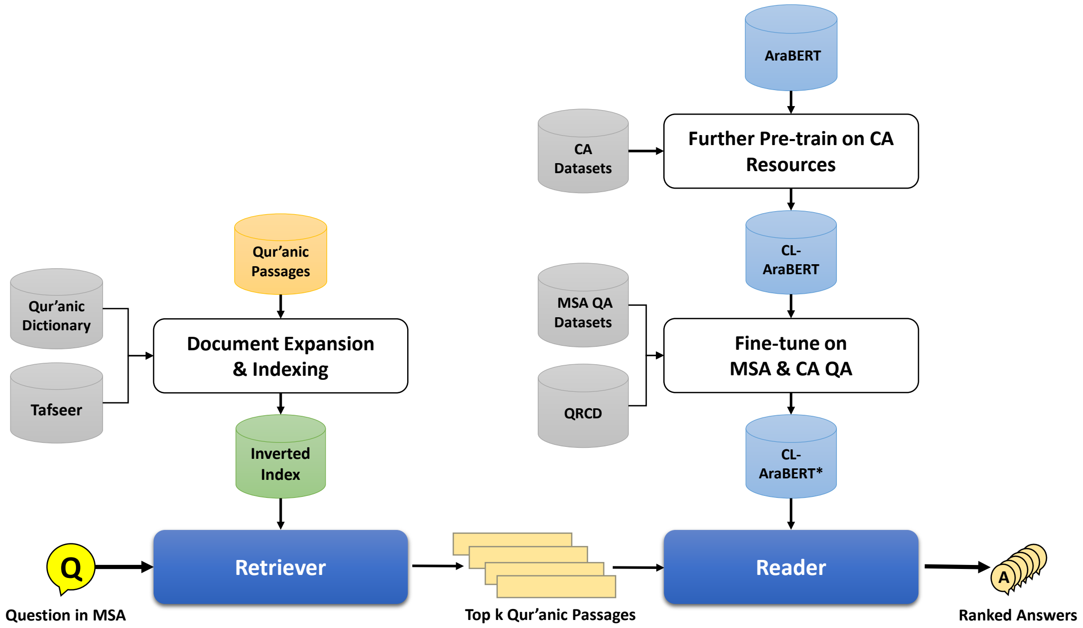


Figure 1.5. An overview of the pipelined Retriever-Reader architecture of the QA System.

The need to evaluate our CL-AraBERT reader and the end-to-end QA system on multi-answer questions was an eyeopener to the absence in the literature of a rank-based measure that can *fairly* integrate partial matching for that type of questions. Although the currently used set-based measures for evaluating extractive QA/MRC systems on multi-answer questions (in the literature [49], [71]) can integrate partial matching, they are not rank-based. Moreover, even with partial matching of answers, there are cases where the matching can be *unfair* specifically when predicted answers comprising more than one gold answer are matched to only one of the best matching gold answers, but not more. To address the aforementioned issues, we introduce a simple yet novel method to match the predicted answers against their respective gold answers; and we adapt the traditional Average Precision (AP) rank-based measure to integrate *partial* matches in addition to exact matches of answers. We denote this measure as *Partial Average Precision* (pAP). For evaluating the CL-AraBERT reader and the QA system, we used pAP for both multi-answer and single-answer questions, in addition to the traditional measures for single-answer questions.

Finally, as a gesture to promote state-of-the-art research on Arabic QA and MRC tasks over the Holy Qur'an, the *QRCD* dataset was used to organize a QA shared task on the Qur'an to stimulate the interest of the research community on the task [87].

1.3. Research Questions

Before introducing the research questions in this section, we formally define the two question types in *QRCD* to motivate the research questions. A *Single-answer* question is the question that has only one answer (i.e., an answer that is a single span of text, denoted as an "answer span") in the accompanying Qur'anic passage, as shown in Figure 1.1-(a) and (c). On the other hand, a *multi-answer* question is the one whose answers are composed of several components (such as *list* or *why* questions) in two or more different answer spans in the accompanying Qur'anic passage, as shown in

Figure 1.1-(b). *QRCD* was used in evaluating the reader component and the end-to-end QA system on their performance over single-answer questions and multi-answer questions, independently. However, the question types and the scope of the evaluation have implications on the proposed evaluation measures in Section 3.2.3.

We define a passage-scope and a Qur'an-scope for the evaluation. The *passage-scope* is confined to the passage accompanying the question to which its answer(s) were extracted from, while the *Qur'an-scope* comprise the whole Qur'an given that the answer(s) to a given question (of type single-answer or multi-answer) may appear in semantically and/or syntactically similar forms in different chapters and across different verses within different Qur'anic contexts.

Based on the forgoing, the passage-scope is adopted for evaluating the reader component, and the Qur'an-scope is adopted for evaluating the retriever component and the end-to-end QA system. However, this implies that a multi-answer question with two or more answer components (i.e., answer spans) in the Qur'an, will be evaluated as a multi-answer question in the Qur'an-scope evaluation of the end-to-end QA system; and it may be evaluated as a single-answer question in the passage-scope evaluation of the reader component, if the question happens to be coupled with a Qur'anic passage comprising only *one* of the question's answer components. As such, the adopted scope will also influence whether the question is classified as single-answer or multi-answer.

In this work, we address six major research questions.

RQ1: Would expanding the Qur'anic passages with their corresponding Qur'an related MSA resources help the retriever in bridging the gap between the questions in MSA and their answer-bearing Qur'anic passages?

RQ2: Since our model is the CA extension of the MSA-only AraBERT, does further pre-training with Classical Arabic improve the performance over the MSA-only pre-trained model?

RQ3: With the relatively-modest size of *QRCD*, would it be enough to exclusively rely on transfer learning from MSA to CA in fine-tuning the readers without the need for MRC datasets in Classical Arabic?

RQ4: Adopting the passage-scope for evaluation, how does the fine-tuned CL-AraBERT reader perform on multi-answer questions vs. single-answer questions?

RQ5: Adopting the the Qur'an-scope for evaluation, how does the end-to-end QA system perform on multi-answer questions vs. single-answer questions?

RQ6: Is a native BERT-based model architecture fine-tuned as an extractive MRC reader sub-optimal for QA and MRC tasks over multi-answer questions?

Our experiments show that a passage retriever over an Okapi BM25 [113] index of Qur'anic passages expanded with two MSA resources significantly outperformed a baseline retriever over an index of Qur'anic passages only. Moreover, we empirically show that the fine-tuned CL-AraBERT reader model significantly outperformed the similarly fine-tuned AraBERT model, which is the baseline. In general, the CL-AraBERT reader performed better on single-answer questions in comparison to multi-answer questions. Moreover, it has also outperformed the baseline over both types of questions. Furthermore, despite the integral contribution of fine-tuning with the MSA datasets in enhancing the performance of the CL-AraBERT and AraBERT readers, relying exclusively on those datasets (without MRC datasets in CA, e.g., *QRCD*) may not be sufficient for our reader models. This finding demonstrates the relatively high impact of the *QRCD* dataset (despite its modest size). As for the QA system, it consistently performed better on single-answer questions in comparison to multi-answer questions. However, our experiments provide enough evidence to suggest that a native BERT-based model architecture fine-tuned on the MRC task may not be intrinsically optimal for multi-answer questions.

1.4. Contributions

Our contribution in this work is nine-fold:

(1) We introduce *AyaTEC*, the first fully-reusable test collection for Arabic question answering on the Holy Qur'an where *all* the qur'anic verses that directly answer the questions were exhaustively extracted and annotated.[6] *AyaTEC* targets the information needs of curious and skeptical users. It is also diverse in its topic categories and covers factoid and non-factoid questions.

(2) We extend *AyaTEC* to introduce *QRCD* as the first extractive machine reading comprehension dataset on the Holy Qur'an.

(3) To facilitate the use of *AyaTEC* and *QRCD* in evaluating Arabic QA and MRC systems on the Holy Qur'an, we propose several evaluation measures to support the different types of questions and the nature of verse-based answers, and span-based answers. We introduce *Partial Average Precision* (pAP) as the rank-based measure that integrates partial matching to evaluate performance over multi-answer as well as single-answer questions. We also introduce a simple yet novel method to fairly match predicted answers of multi-answer questions against their respective gold answers.

[6] A user of our test collection detecting the absence of a verse (or set of verses) that potentially answers a question directly or indirectly is urged to contact the authors.

(4) We demonstrate the effective contribution of expanding the Qur'anic passages with corresponding MSA resources, in assisting the retriever to mitigate the gap between the questions in MSA and their answer-bearing Qur'anic passages.

(5) We introduce CL-AraBERT, which is a further pre-trained version of the AraBERT model [23], using a large Classical Arabic dataset. We then fine-tune it as a reader over Qur'anic passages, before integrating it into a pipelined retriever-reader architecture to constitute (to the best of our knowledge) the first extractive MRS QA system on the Holy Qur'an.

(6) We demonstrate the integral contribution of cross-lingual transfer learning from MSA to CA, by empirically showing that it is essential to complement MSA resources with CA resources to attain better performance on the reading comprehension task on the Holy Qur'an.

(7) We empirically provide enough evidence to suggest that a native BERT-based model architecture fine-tuned on the MRC task may not be intrinsically optimal for multi-answer questions.

(8) We make the pre-trained CL-AraBERT model, the *AyaTEC* dataset,[7] the *QRCD* dataset set, and the evaluation script publicly available to promote state-of-the-art research on QA and MRC tasks over the Holy Qur'an.[8]

(9) Hoping to trigger state-of-the-art research on Arabic QA and MRC tasks over the Holy Qur'an, the *QRCD* dataset was used to organize a QA shared task on the Qur'an to stimulate the interest of the research community. Thus, forming a seed for growing a virtual research community on Qur'anic research.

[7] *AyaTEC* can be downloaded from `http://qufaculty.qu.edu.qa/telsayed/datasets`

[8] All (except *AyaTEC*) can be downloaded from this link `https://github.com/RanaMalhas/QRCD`

The rest of this dissertation is organized as follows. We cover the literature review in Chapter 2, and dedicate Chapter 3 to describe our methodology in building the two Qur'anic QA datasets (*AyaTEC* and *QRCD*) with an evaluation perspective. In Chapters 4 and 5, we cover the development and evaluation of the *Retriever* and *Reader* components of the *Retriever-Reader* architecture of our QA system, respectively. Then in Chapter 6, we describe our methodology in integrating the *Retriever* and *Reader* components into a pipelined *Retriever-Reader* architecture, to constitute our end-to-end *machine reading at scale* QA system on the Holy Qur'an. We conclude that chapter with general implications of our research work. Then we conclude this dissertation with a summary of the main findings and contributions of this work, before presenting our thoughts towards future work.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we review existing Qura'nic Arabic QA datasets with an evaluation perspective, and then discuss the potential of using these datasets in MRC tasks if they are to be extended (Section 2.1). Then we cover existing Arabic QA systems and search tools on the Holy Qur'an (Section 2.2). With the resurgence of MRC as an integral component of modern QA systems, we overview existing Arabic reading comprehension datasets and systems (Section 2.3) before discussing important transformer-based MRC models in the literature (Section 2.4). We conclude this chapter with an overview of the main approaches adopted by the participating teams in the QA shared task that we have organized on the Holy Qur'an [87].

2.1. Arabic QA Datasets on the Holy Qur'an

In this section, we shed light on the main performance evaluation methodology adopted by prominent work on Arabic question answering on the Holy Qur'an in the literature, and review existing Arabic QA datasets on the Qur'an.

2.1.1. Evaluation of Arabic QA on the Holy Qur'an

Abdelnasser, Ragab, Mohamed, *et al.* [5] evaluated their overall QA system by five Qur'an experts using 59 test questions that were not made publicly available. Hamdelsayed and Atwell [58] and Hamdelsayed, Mohamed, Saeed, *et al.* [59] used 30 questions from the QA test collection developed by Hamdelsayed and Atwell [57], but they resorted to Islamic scholars to evaluate their respective systems. Similarly, Hakkoum and Raghay [56] evaluated their QA system using 52 test questions that were developed by an Islamic studies researcher, and the retrieved answers were manually judged. The questions were made available but without their answers. Hamdelsayed and Atwell [58], Shmeisani, Tartir, Al-Na'ssaan, *et al.* [123], Ouda [102], and Hamoud and Atwell [61] also adopted a similar evaluation approach. This overview implies that evaluation of Arabic QA research based on Qur'an experts' judgement of systems' returned answers does not warrant fair performance comparisons due to the use of different sets of questions.

2.1.2. Existing Arabic QA Datasets on the Holy Qur'an

In this section we overview the few existing Arabic Qur'anic QA datasets, and discuss their potential for use in evaluating and/or training QA and MRC systems. To the best of our knowledge, there are no *extractive* MRC datasets on the Holy Qur'an in the literature. *Extractive* MRC refers to the task of span prediction, where the answer is a specific span of text extracted from passage(s) accompanying a question. Whereas *generative (abstractive)* MRC refers to the task of answer generation, where the answer is formulated using natural language, and is not necessarily confined to a span of text [34], [38], [148]. For a dataset to be suitable for use in MRC tasks, it should comprise question-answer pairs that are coupled with the passage(s) or document to which the answers were extracted/generated from (to form tuples of question-passage-answer triplets) [38].

There are three relatively recent Arabic Qur'anic QA datasets (test collections) [16], [57], [62] that can be used as training resources on the QA task, but have some limitations towards their reusability in evaluation as explained below.

The QA dataset of Hamdelsayed and Atwell [57] is composed of 263 Arabic questions with a total of 263 question-answer pairs. Similar to *AyaTEC*, the gold standard answers to these questions are qur'anic verses with each answer constituting one verse or a set of consecutive verses. The answers were validated and judged by Islamic scholars, and no information was shared about the question types. Although Hamdelsayed and Atwell's test collection has potential because it is verse-based, it is not fully-reusable because the questions and answers are drawn only from the first two chapters of the Holy book; namely, Al Fatiha and Al Baqara. This would limit its usability as a test collection by QA systems targeting the whole Qur'an since a relevant answer (qur'anic verse/verses) may be repeated in chapters other than the first two.

The QAEQ&AC (Qur'an Arabic-English Question and Answer Corpus) by Hamoud and Atwell [62] is another potential dataset that is composed of 1500 question-answer pairs, of which 1000 are Arabic and 500 are English. They were developed or acquired using four different sources including: FAQs from well known Islamic QA forums, manually devised questions and drawn answers from the Qur'an, questions of some Muslims in the Holy Mosque in Mekka answered by attending Islamic scholars, and test sets from previous QA research work. The answers are mainly in natural language text with some being qur'anic verses. Hamoud and Atwell did not release information about the distribution, coverage or topic diversity of the QA pairs. Although a lot of effort was invested on extracting and cleaning the data, it was not mentioned if the final answers to the questions were validated by Qur'an scholars or specialists, especially those taken from test sets used by earlier published work. Nevertheless, Hamoud and Atwell have mentioned that the dataset/test collection will not be made publicly available until all the answers are validated by Qur'an scholars. Given the aforementioned profile of the QAEQ&AC dataset, it would not be fully-reusable in evaluating QA systems for the Qur'an, since it does not include an exhaustive set of all relevant answers to the respective questions from the Holy book.

The AQQAC (Annotated Corpus of Arabic Al-Qur'an Question and Answer) by Alqahtani and Atwell [16] and Alqahtani [17] is a notable QA dataset with 2224 QA pairs, of which 1000 were extracted from a book by Ashur [26]. The remaining 1224 QA pairs were scrapped from a website on Al-Qur'an and Tafseer,[1] whose answers were extracted from Tafseer Al-Tabari [128]. The aforementioned two sources are considered trusted Islamic resources. Due to copyright concerns, the publicly available part of the AQQAC dataset only comprise the 1224 QA pairs. Nevertheless, it is considered an important resource for training QA systems on the Qur'an. However, although the questions cover the whole Qur'an, their answers are not exhaustive; i.e., not all the Qur'anic verses that answer a given question were exhaustively extracted from the whole Qura'n. As such, the AQQAC dataset has the limitation of not being fully-reusable in evaluating QA systems on the Qur'an.

Based on the foregoing, *AyaTEC* [85] has been designed to fill the above identified gap and to address the limitations of using existing QA datasets/test collections for the Qur'an in evaluation. *AyaTEC* is fully-reusable by including, to the best of effort, an exhaustive set of all relevant/direct answers (qur'anic verses) to the questions that may be repeated (in different contexts) in any of the 6236 verses of the Holy Qur'an. Moreover, several quality measures were adopted to ensure a reliable judging/evaluation of the answers' relevance to the questions. Such measures include seeking three specialists in Holy Qur'an Interpretation (Tafseer), who have completely memorized the Qur'an, for the judging task. Moreover, the Fleiss kappa [124] statistic was used as an indicator of inter-rater agreement among the judges. Furthermore, none of the former QA test collections have tackled the issue of integrating partial matching into the measures used in evaluating QA systems. Partial matching integration is important and inherently poses

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

itself as a necessity, since answers may include one or a set of consecutive qur'anic verses that may or may not be all retrieved by systems.

Potential for using the QA datasets as MRC datasets

On the MRC front, the published part of the AQQAC dataset (1224 QA pairs) [16] is the easiest to extend and transform for use as a *generative* MRC dataset because the answer to each question is composed of two parts, a natural language answer and its related verse-based answer. The latter may serve as a context (passage) to the former, if the verse-based answer is long enough, which may not be the case since the majority of answers in AQQAC are composed of only one verse. Augmenting single verse-based answers with neighboring verses would be a possible solution to constitute Qur'anic passages. Naturally, the AQQAC cannot be readily used as an *extractive* MRC dataset, unless the exact answer spans of text are extracted from their corresponding contexts.

As for *AyaTEC*, its verse-based answers have also served as contexts to the answers that were extracted by the annotators to develop *QRCD* as an *extractive* MRC dataset (as we describe in Section 3.2). Though thematic passage curation was adopted using the Thematic Holy Qur'an[2] [127] to segment the Qur'anic text into passages that served as larger contexts for the extracted answers (as detailed in Section 3.2.1.1), to make it a better fit for the task. With respect to evaluation, unlike AQQAC, *QRCD* can be used for evaluating MRC and QA systems since it is fully-reusable, as the answer spans for each question were extracted from the exhaustive direct answers in *AyaTEC*; i.e. all answer spans that may answer a given question, were extracted from the whole Qur'an (to the best of effort).

Similarly, the dataset by Hamdelsayed and Atwell [57] also has the potential to be

[2] `https://surahquran.com/tafseel-quran.html`

extended and transformed into an MRC dataset using a similar approach to that used for developing *QRCD*. Whereas transforming QAEQ&AC [62] is much more challenging given that not all answers are verse-based.

2.2. Arabic QA and Search Systems on the Holy Qur'an

Alqahtani and Atwell [15] has classified existing search work (including QA work) on the Holy Qur'an into *semantic-based* and *keyword-based* (or text-based) approaches. *Semantic-based* approaches are concept-based relying heavily on ontologies and/or a knowledge base, while *keyword-based* approaches rely mainly on term overlap (i.e. keyword matching). This classification is not mutually exclusive due to the propensity of some approaches to adopt a hybrid of both. We adopt this classification for the review in the next two sections, noting that we also cover and classify embedding-based (or dense) approaches under the semantic-based category.

Both approaches have their own set of limitations on the Qur'an. Keyword search approaches tend to retrieve irrelevant verses or miss to retrieve all relevant verses, especially those that are semantically similar to the query/question, but with minimal term overlap (vocabulary mismatch problem). Whereas, semantic search approaches are predominantly ontology-based; they either suffer from using ontologies that do not cover all the concepts in the Qur'an, or they use more than one ontology (to enhance concept coverage) at the expense of attempting to align the different concept representations among the ontologies they deploy. As such, these semantic search/QA approaches tend to miss retrieving all semantically relevant verses to a query, or miss answering questions on concepts not well represented in the ontology or ontologies they deploy [17], [35]. Moreover, the majority of semantic and concept based approaches do not satisfy the information needs of users because they are designed from a topic-oriented perspective rather than a user-oriented perspective [35].

As our proposed QA system adopts a modern pipelined retriever-reader architecture, our review of existing Arabic Qur'anic QA systems may reformulate their functionality in a modular way to facilitate comparison with our sparse retriever component and answer extraction (i.e., reader) component. Traditional QA systems may also include a question analysis component that is typically responsible for question classification and/or query reformulation. With the close affinity between reading comprehension and question answering, and the fact that the Holy Qur'an is a closed text corpus of numbered chapters and verses, our review also explores whether the answer extraction components of the Arabic QA systems on Qur'an can be perceived as *generative* MRC components as opposed to *extractive* ones.

2.2.1. Existing Arabic QA Systems on the Holy Qur'an

In this section, we review existing keyword-based [61], and semantic-based [6], [55], [102], [123] Arabic QA systems on the Holy Qur'an. We conclude this section with some perceptions towards semantic ontology-based approaches, in addition to some prospects towards enhancing our proposed QA system.

Hamoud and Atwell [61] developed a simple keyword-based QA system over their QAEQ&AC corpus [62]. Given a factoid or non-factoid question, the retriever simply uses regular expressions to retrieve questions from the QAEQ&AC corpus that have high term overlap with the terms of the given question. The retrieved questions are then re-ranked using a keyword-based question-question similarity scoring function that is also based on term overlap. The answer of the top scoring question is returned as the answer to the given question. Hence, the system does not have an answer extraction component.

Hakkoum and Raghay [55] developed a QA system powered by a semantic-based search engine (as the retriever component) that leverages a Qur'anic ontology they built to represent the knowledge and concepts of the Qur'an in Web Ontology Language format. Given a question in MSA, a natural language interface (NLI) reformulates the question into a SPARQL query (Protocol and RDF Query Language- the standard query language for the Semantic Web), which the retriever uses to retrieve the candidate answers to that query from the Qur'anic ontology. If no match is found, query expansion is adopted as a rescue. The system does not have an explicit answer extraction component because it mainly relies on the query generation component to achieve better answer retrieval. Shmeisani, Tartir, Al-Na'ssaan, *et al.* [123] also adopted a semantic-based approach for their QA system that is highly similar to that of Hakkoum and Raghay [55], but it was only applied on factoid questions. Both QA systems do not have an answer extraction component.

Abdelnasser, Ragab, Mohamed, *et al.* [6] developed a semantic-based QA system (Al-Bayan) that is composed of a question classifier, a retriever and an answer extraction component. A SVM classifier classifies a given question posed in MSA into a taxonomy of answer types or NER (Named Entity Recognition) classes. Then a semantic-based retriever attempts to match the question to a concept in their built Qur'anic ontology, and retrieves all relevant verses and their respective interpretations as candidate answers. The answer extraction component ranks those candidate answers using a NER model and a set of text-based features. If the system fails to match a question to a concept in the Quranic ontology, no answer is returned. A limitation of this system is its design to answer factoid questions only.

Ouda [102] developed a multi-purpose system (QuranAnalysis) which includes a question answering module that accepts a question in Arabic or English (the case of English question is not covered in this review). QuranAnalysis also adopts a semantic ontology-based approach. It is composed of three components; a question analysis component, followed by a semantic retriever and answer extraction components. Given a question, the question analysis component conducts a form of query expansion by enriching the question with all possible derivations and synonyms of its key words. Then, the semantic retriever tries to match the terms in each question with all relevant concepts in the ontology using a similarity score. This score is computed using a minimum edit distance and a character similarity algorithm. Finally, the answer extraction component extracts candidate answers from the ontology in two ways depending on whether the question term is a concept in the ontology or not; if the question term is a concept, then all inbound relations with that concept are retrieved as objects to formulate candidate answers. Whereas, if the question term is not a concept, then all relation verbs associated with the question term are extracted as potential answers (assuming in this case that an answer is a verb). Moreover, candidate answers were also extracted from the Qur'an verses using question-verse similarity. Eventually, all candidate extracted answers were sorted based on their similarity scores and the top answer was returned.

We believe that the answer extraction methodology adopted in [6], [102] could be perceived as a form of generative MRC, given that the Holy Qur'an is a closed text corpus of numbered chapters and verses with its knowledge represented as concepts in ontologies. Such Qur'anic ontologies may also comprise links to the actual verses and their interpretations (such as the case in [6]). However, there are several limitations to ontology-based approaches. Many of the developed ontologies do not cover all the concepts in the Qur'an, and they may adopt different taxonomies for the Qur'an concepts/topics. Thus, making the task of merging ontologies a very challenging one, as reported in [6]. Moreover, QA and search systems adopting this approach are, by design, concept or topic oriented rather than user oriented. As such, they may not be optimal in addressing the information needs of users seeking specific answers to their questions and queries [35].

In contrast, our proposed QA system is designed to handle factoid and non-factoid questions, and to address the information needs of curious and skeptical users. This is attempted through the extractive MRC reader that we fine-tune using the *QRCD* dataset, which comprise questions raised by the two user types. However, despite adopting a semantic-based passage expansion approach in our sparse (keyword-based) retriever component, our system would very well benefit from adopting dense (embedding-based) retrieval approaches that may be better at capturing semantically relevant Qur'anic passages. We elaborate such prospects in Section 7.2. In general, we believe that a hybrid of keyword-based and semantic-based (embedding and/or ontology-based) approaches can be integrated for better performance.

2.2.2. Existing Arabic Search Systems on the Holy Qur'an

Unlike Arabic Qur'anic QA research work, Arabic search systems and research on the Holy Qur'an is more prevalent because it is a relatively active area of research. Although not covered in this review, we point out that there is a wide presence of Qur'anic web[3] and mobile applications[4] with search tools that largely adopt keyword-based approaches, with some adopting semantic and concept-based search approaches[5] [95], but with a lower extent.

Acknowledging the limitations of keyword search, Arabic Qur'anic search research includes more presence of keyword-based approaches that are enhanced with semantic-based query expansion approaches, in addition to semantic-based approaches or hybrids of semantic-based and keyword-based approaches. In the remainder of this section, we review notable papers from each category, then conclude with remarks to position our proposed retriever component with respect to similar search work on the Qur'an.

2.2.2.1. Keyword Search Approaches

Early Arabic search systems on the Holy Qur'an explored the effect of query expansion when coupled with keyword search over an inverted index. Hammo, Sleit, and El-Haj [60] showed that expanding query words with their respective synonyms using a Thesaurus has warranted an improvement in the performance of their index search over the verses of the Qur'an. The Thesaurus they have used was developed by grouping Qur'anic words into semantic word classes for the purpose of query expansion. Three inverted indexes were used; a vowelized-word index of distinct Qur'anic words with their diacritics, a non-vowelized-word index of normalized distinct Qur'anic words, and

[3] Prominent web applications with keyword-based search tools include *Tanzil* `http://tanzil.net`, *KSU Digital Mushaf* `http://quran.ksu.edu.sa/`, *Al-Munaqib Al-Qur'any* `http://www.holyquran.net/search/sindex.php` among many others.

[4] Mobile apps with keyword-based search tools include *Ayat* `https://play.google.com/store/apps/details?id=sa.edu.ksu.Ayat`, *DiamondQuran* `https://play.google.com/store/apps/details?id=com.DiamondQuran` among many others.

[5] Semantic and concept-based search applications include *Quran by Subject* `https://play.google.com/store/apps/details?id=com.Quran1.hello`, and *Holy Quran Search Engine* `https://play.google.com/store/apps/details?id=com.fara.quransearch` among others.

a root-based index. Al Gharaibeh, Al Taani, and Alsmadi [12] adopted a similar index search approach with query expansion using synonyms of the Microsoft Word Arabic Thesaurus. Naturally, this resource was not very effective in enhancing performance due to its poor coverage of Qur'anic Classical Arabic words. Yusuf, Yunus, Wahid, *et al.* [145] also adopted query expansion to enhance their keyword-based index search, but they used a two-phase query expansion technique. They first use lexically similar words to expand the query, then find their corresponding contextually related words in a Qur'an ontology that captures words' relationships to further expand the query. As such, they have used similar and related words in query expansion. Beirade, Azzoune, and Zegour [36] adopted a similar approach to [60] by developing a search engine using a lucene inverted index and building an ontology of Qur'an words with the semantic relations among them for use in query expansion.

2.2.2.2. Semantic Search Approaches

On the semantic search front, research on the knowledge representation of the concepts and topics of the Holy Qur'an in Arabic has taken its fair share in the literature, with ontology-based representations dominating. As such, the majority of semantic search approaches are ontology-based in which search is facilitated through constructing a structured query (such as SPARQL) from natural language queries to retrieve relevant verses. Examples of such semantic search approaches include the work of Sherif and Ngomo [122] and Yauri, Kadir, Azman, *et al.* [143]. On the other hand, Alhawarat [13] adopted a Latent Dirichlet Allocation (LDA) topic modeling approach to represent the topics of Chapter 12 (Surat Joseph (PBUH)) in the Qur'an. Their experiments provided enough evidence to suggest that LDA topic modeling may not be effective when applied on the Qur'an (given the very modest results attained by the model when applied on the topics of Chapter 12 of the Qur'an). A relatively recent and notable semantic search approach on the Qur'an was introduced by Mohamed and Shokry [95], in which they develop an embedding-based search tool. They first trained a continuous bag-of-words (CBOW) word2vec model [92] using two large Classical Arabic datasets [18], in addition to three MSA resources that include two news datasets (BBC-Arabic and CNN-Arabic) [115], and a dataset of Arabic book reviews [22]. Then, they used the Qur'an concepts taxonomy of "Mushaf Al Tajweed" [54] (similar to [2]) to manually annotate each verse in the Qur'an dataset. For search, the trained word2vec model was used to generate feature vectors for the query and the topics in the taxonomy, each represented by the words it comprises. For retrieval, the cosine similarity (dot product) between the query vector and each of the topic vectors was used to retrieve the most topic-relevant verses to the query.

2.2.2.3. Hybrid Search Approaches

With respect to hybrid search approaches that harness the benefits of keyword and semantic search paradigms, Abbas [2] developed a bilingual (Arabic and English) search tool over Qur'anic concepts. It is composed of two modules; a keyword search module and a tree of concepts module. To increase the effectiveness of the keyword search module, query expansion using eight English translations of the Holy Quran in addition to the Arabic version were used. To further enhance the keyword search tool, it was integrated with a tree of Qur'anic abstract concepts built based on the categorization of "Mushaf Al Tajweed" [54]. Alqahtani and Atwell [14] also proposed a search system that is a hybrid of both search paradigms. The system first tries to match a given query to a concept in their built Arabic-English Qur'an ontology to retrieve all relevant verses. If the query does not match any concept, keyword search is adopted using word matching over an index. They developed their ontology by aligning several Qur'an ontologies in an attempt to cover all the concepts in the Qur'an. Safee, Saudi, Pitchay, *et al.* [117] proposed a similar hybrid search approach. Their semantic search was conducted over an ontology that they have developed to only cover the medical and health science knowledge in Qur'an. Recently, Zouaoui and Rezeg [149] have adopted a hybrid search approach that attempts to overcome the vast number of verses returned by ontology-based search engines in response to a user query, which is considered a known limitation to these hybrid approaches. To this effect, they adopted Earab (i.e., Arabic grammar rules) in a novel method to construct a Qur'an ontology as an index. They emphasized the importance of deriving and representing the relations between the words of each Qur'anic verse. Such word relations are exploited and applied on the query words at search time to enhance verse retrieval. Given the sacredness and sensitivity of the Qur'an's content, they used a semi-automatic method to construct the ontology with the intervention of Qur'an scholars.

Our review of the prevalent Arabic Qur'anic search approaches in the literature, has revealed that our retriever component (of the proposed QA system in this work), is the first keyword-based Arabic Qur'anic search system to adopt document (i.e., passage) expansion rather than query expansion. Nevertheless, many of the semantic ontology-based search approaches did integrate Qur'an-related resources in their respective Qur'an ontologies. We iterate that our retriever component can benefit from integrating some form of semantic search technologies, either through dense embedding-based retrieval approaches or ontology-based ones, to complement the strengths and overcome the limitations of keyword and semantic search paradigms. Further prospects towards enhancing our retriever component are elaborated in Section 7.2.

2.3. Existing Arabic Reading Comprehension Datasets and Systems

On the MSA front, we overview notable datasets and systems with emphasis on those that were landmarks in influencing the progress of Arabic reading comprehension systems in the literature. The QArabPro [11] is a rule based reading comprehension system that was evaluated on a dataset of 335 factoid and non-factoid questions over 75 reading comprehension tests. In 2012 and 2013, the Question Answering for Machine Reading (QA4MRE) task was organised at the CLEF (Cross-Language Evaluation Forum) for several languages with Arabic being one of them [105]. The QA4MRE datasets at CLEF 2012 and CLEF 2013 were composed of 160 and 240 multiple choice questions, respectively, coupled with their 16 accompanying test documents. IDRAAQ [8] and ALQASIM [52] were among the participating systems in CLEF 2012 and CLEF 2013, respectively. IDRAAQ heavily relied on its passage retrieval (PR) module to answer the questions. ALQASIM adopted a new approach (back then) by first analyzing the reading test document, then analyzing the questions and each of their corresponding multiple choice answers before selecting an answer. Another interesting comprehension approach that is based on Rhetorical Structure Theory (RST) [90] was proposed by Azmi and Alshenaifi [31] in their LEMAZA QA system, to answer Arabic *why* questions. Discourse analysis was used to identify cue phrases (i.e., words and phrases that serve as unit connectors), which they leverage to build the rhetorical relations between textual units. A candidate answer-bearing passage to a given question is represented using their

RST method before extracting and generating the candidate answer(s) to the question from this passage [21]. LEMAZA was evaluated using 110 *why* questions over a dataset of 700 articles extracted from the OSAC Arabic corpus [115]. Other non-traditional reading comprehension approaches include those based on textual entailment between the logical representation of a given factoid question and the passage to which an answer is extracted from [21], [32], [33]. Starting from 2018 onwards, relatively larger Arabic MRC datasets started to appear in the literature. Ismail and Homsi [65] developed their DAWQAS dataset, which is composed of 3025 question-passage-answer triplets for *why* questions that were scraped from Arabic websites.

The next two MRC datasets to overview are those developed by Mozannar, Maamary, El Hajal, *et al.* [97]. The two datasets (combined) have marked the beginning of Arabic neural reading comprehension models. The first is the Arabic Reading Comprehension Dataset (ARCD) which is composed of 1,395 question-passage-answer triplets whose questions were generated by crowdsource workers from their accompanying contexts of Arabic Wikipedia passages. The second is the Arabic SQuAD, which is the Arabic translated version of the English SQuAD v1.1. It comprises 48.3k QA pairs translated with their corresponding articles. Only factoid questions were included. Mozannar, Maamary, El Hajal, *et al.* developed SOQAL, which is a system for open-domain QA for the Arabic language that adopts the retriever-reader QA model proposed by Chen, Fisch, Weston, *et al.* [39]. It is composed of a TF-IDF document retriever and a fine-tuned multilingual BERT [48] reader over Wikipedia articles. Both datasets were used in fine-tuning the MRC reader of their SOQAL system. It was not long before the release of AraBERT [23] and later AraELECTRA [24], which are the Arabic versions of BERT and ELECTRA [43], respectively. The two datasets by Mozannar, Maamary, El Hajal, *et al.* were also used in fine-tuning AraBERT and AraELECTRA as reader models.

Another MRC dataset with a relatively large size is the AQAD dataset [27]. It is composed of about 17k QA pairs for 3,381 passages extracted from 299 Arabic wikipedia articles. The selected Arabic articles correspond to a set of English wikipedia articles in the SQuAD dataset. The corresponding factoid questions of those selected SQuAD articles were translated to Arabic using Google Translate. The AQAD dataset was used in fine-tuning a multilingual BERT model and a BiDAF (Bidirectional Attention Flow for Machine Comprehension) model [120] as MRC readers. The last datasets to overview are two multilingual MRC datasets, each having a fair share of Arabic questions. The TyDi QA [42] and MLQA [79] datasets comprise 26K and 5k Arabic questions, respectively. The main purpose of developing these datasets is to conduct extensive transfer learning QA experiments across languages (including Arabic) using different training/testing settings, including zero-shot transfer. The datasets were used in fine-tuning pre-trained multilingual and mono-lingual BERT-based language models as cross-lingual MRC readers. Naturally, the Arabic portions of these datasets can be exploited in fine-tuning mono-lingual Arabic transformer-based MRC readers as well.

Our adopted extractive MRC approach in this paper is inspired by AraBERT. Our work extends AraBERT by further pre-training the MSA-only pre-trained model using Classical Arabic, to make it a better fit for our MRC task on the Holy Qur'an. We consider our task more challenging because the system needs to answer non-factoid (and factoid) questions with one or more answers, as opposed to only factoid questions with only one answer. Among the overviewed MSA datasets, only two datasets include questions with more than one answer; namely, the dataset used in evaluating LEMAZA [31] and the DAWQAS [65] dataset. The LEMAZA system handled multi-answer questions by returning the answer with the highest priority for its RST relation. Though, it can be extended to return all answers to a multi-answer question ranked by their RST priority scores. As for the DAWQAS dataset, no baseline or QA system was reported to have used this dataset. This makes our Arabic MRC system among the few that have catered for answering multi-answer questions.

2.4. Machine Reading Comprehension

MRC has been recently fueled by the success of transformer-based [129] pre-trained language models, exemplified by the phenomenal success of BERT [48] and BERT-like models [43], [83] on answer extraction tasks over MRC datasets, such as SQuAD. As our approach is BERT-based, we overview other important transformer-based models and architectures that we may adapt in future work using the same CA resources that we have developed and used in this work.

In general, what makes pre-trained language models very appealing is their unsupervised transfer learning potential, and generic architectures that can be minimally adapted to work for several different downstream NLP tasks (including MRC), by simply fine-tuning an additional task-specific output layer on relatively small sized labeled data. The advent of BERT in 2018 marked a new era for NLP; its bidirectional encoder-only transformer for text representation gained its competitive edge over its rivals (at that time [106], [109]), by jointly attending and conditioning on left and right contexts across all transformer layers. It was not long before the inception of a fleet of BERT descendants and peers (with encoder-only, decoder-only, or encoder-decoder transformer architectures) that outperformed BERT on many NLP tasks. Some of the most prominent post-BERT models that performed well on the reading comprehension task include XLNet [141], RoBERTa [83], GPT-3 [37], ELECTRA [43], BART [78], SpanBERT [66], DeBERTa [63], InstructGPT [103] and its more recent successor ChatGPT[6] among others. We intentionally leave out describing these models except for SpanBERT, because it is inherently suitable for the span prediction task due to its span-masking (rather than token-masking) scheme. The model is pre-trained to predict the masked spans using span-boundary representations and a span-boundary objective [66].

Despite the success of the above extractive MRC transformer-based approaches on single-answer questions, only few of them focused on *multi-answer* questions that require reasoning over multiple sentences.[7] This is mainly attributed to the scarcity of large English datasets with multi-answer questions for extractive MRC. Current datasets that we came across include: MultiRC [71], DROP [49], QUOREF [47], and WikiHowQA [44]. Many transformer-based models that were fine-tuned using these datasets achieved satisfactory performance despite being initially designed for single-answer questions; e.g., RoBERTa, BERT, XLNet and QANet [144], among others. However, other recent MRC approaches have appeared that are specifically designed for multi-answer questions, which outperformed the former models on this task. Dua, Wang, Dasigi, *et al.* [49] and Hu, Peng, Huang, *et al.* [64] employed *multi-head architecture* models on the DROP and QUOREF datasets, respectively. Each head is responsible for predicting an answer span. The number of needed prediction heads is either pre-specified or dynamically predicted and allocated depending on the question type (and its expected answer type). Moreover, Segal, Efrat, Shoham, *et al.* [119] proposed an approach that casts the extractive multi-span prediction problem as a sequence tagging task, in

[6] https://openai.com/blog/chatgpt/

[7] There are MRC approaches that require multi-sentence reasoning to answer *single-answer* questions, such as [112], [139].

which they employ a transformer-based model like BERT for encoding contextualized representations of input question-passage pairs and start/end tokens of each answer span. Their model outperformed former models on the DROP and QUOREF datasets. Finally, ListReader, is a more recent multi-span prediction model proposed by Cui, Hu, and Hu [44] that was trained on their introduced English WikiHowQA dataset.[8] ListReader employs a sequence tagging module that is preceded by an interaction layer composed of a graph neural network, which has two modules. The first module aligns the given question-passage pair to capture relevance, while the second captures inter-answer dependencies among the answer spans in the given passage. Evaluating ListReader on the WikiHowQA benchmark showed that it significantly outperformed the former three models [49], [64], [119] on the same benchmark.

The above overview is an eye-opener to the need for large sized Arabic MRC datasets with multi-answer questions. This is highly needed to facilitate exploiting the above approaches and to advance the development of multi-span extractive MRC models in MSA and Qur'anic Classical Arabic. Except for the moderately sized DAWQAS dataset and the modestly sized QRCD and LEMAZA datasets, all the existing large Arabic MRC datasets (overviewed in Section 2.3) are more adequate for single-span extractive MRC.

2.5. Overview of Systems Participating in the First Shared Task on Question Answering on the Holy Qur'an

Motivated by the resurgence of the machine reading comprehension research, we have used *QRCD* to organize the first Qur'an Question Answering shared task,"Qur'an

[8] Cui, Hu, and Hu [44] also applied ListReader on their introduced Chineze WebQA dataset.

QA 2022" [87]. The task in its first year aims to promote state-of-the-art research on Arabic QA in general and MRC in particular on the Holy Qur'an. First, we briefly describe the shared task that succeeded in attracting 13 teams to participate in the final phase, with a total of 30 submitted runs. Then we outline the main approaches adopted by the participating teams in the context of highlighting some of our perceptions and general trends that characterize the participating systems and their submitted runs.

The shared task definition is similar to the second sub-problem statement (i.e., MRC task) in Section 1.1, but it was relatively simplified such that a system may find *any* correct answer from the accompanying passage (rather than *all* answers), even if the question has more than one answer in the given passage. We also note that the adopted evaluation is different than that adopted in this dissertation in two ways: (i) the test dataset used is a subset of that used for evaluating the MRC reader in Chapter 5, and (ii) the main evaluation measure used in the shared task is Partial Reciprocal Rank (pRR) (defined in Section 3.1.6) as opposed to Partial Average Precision (pAP) (defined in Section 3.2.3).

Pre-training transformer-based Language models trends. As expected, all of the systems of the submitted runs leveraged variants of pre-trained transformer-based language models (LMs), with the majority using an encoder-only BERT-based model architecture. Top performing systems used AraBERT [23] and AraELECTRA [24]. In contrast, only Mellah, Touahri, Kaddari, *et al.* [91] used a multilingual T5 (or mT5) encoder-decoder model architecture [138]. Although such an architecture intrinsically supports sequence-to-sequence generative rather than extractive QA and MRC tasks, their best performing run attained a pRR score that is very close to the median of all pRR scores attained by the 30 submitted runs [87]. Henceforth, any subsequent reference to the median in this section pertains to the median of the pRR scores of these 30 submitted runs.

Naturally, the Arabic language was the main constituent of the dataset(s) used in pre-training the models, 20 of which were pre-trained using MSA-only resources, and the remaining 10 were pre-trained using either multilingual resources, CA-only resources, or a mix of MSA, CA, and dialectal Arabic (DA) resources. Surprisingly, *none* of the LMs pre-trained using CA resources exclusively [69] or partially (using CA as well as MSA and DA resources) [108] have their respective systems/runs achieve above median pRR scores. This is counter-indicative given that the Qur'an is in Classical Arabic. We speculate that adopting pre-trained models using CA-only resources or CA-resources combined with DA resources would prohibit or impede chances of transfer learning from MSA to CA. Albeit, this is needed given that the questions are in MSA and the answers are in CA. In fact, the second research question in this dissertation has tackled this issue. Our findings in Section 5.2.2.1 suggest that classical models pre-trained using MSA and CA resources outperform non-Classical models that are pre-trained using MSA-only resources. Further research is needed to verify the presumably negative effect of pre-training using DA resources alongside CA and MSA resources for QA/MRC tasks over the Holy Qur'an.

Interestingly, only 3 out of the 30 systems further pre-trained their language models using CA resources in an attempt to make them a better fit for the Qur'an QA task. One of the teams (the Rootroo team [87]) further pre-trained two multilingual BERT (mBERT) models [48] using their crawled large corpus of Islamic and Fatwa websites, in addition to the verses of the Holy Qur'an. Whereas Wasfey, Elrefai, Marwa, *et al.* [136] further pre-trained an AraBERT model using only the verses of Qur'an for their relatively least performing run (among their other two better performing runs). This is not expected to make a significant improvement due to the relatively modest size of the Holy Qur'an to be used as the only CA resource in pre-training. Also, the performance of the submitted runs of the Rootroo team remained below the median of pRR scores, which may question the feasibility of further pre-training multilingual rather than monolingual MSA-only pre-trained models. This is a path worth further exploring.

Fine-tuning pre-trained language models trends. With respect to fine-tuning, all the systems used the *QRCD* training dataset in fine-tuning their respective pre-trained language models, either exclusively or in a pipelined fine-tuning procedure, where other (mainly Arabic) MRC datasets were used in fine-tuning prior to using *QRCD*. This is similar to our fine-tuning procedure described in Section 5.1.3. Out of the 30 runs, 10 belonged to systems that only used *QRCD* in fine-tuning [19], [51], [69], [91], [126]. Except for the three runs by ElKomy and Sarhan [51] that leveraged variant combinations of effective post-processing schemes to improve predicted answers, none of the remaining 7 runs attained above-median pRR scores. We speculate that the excelling results of these three systems/runs may have out shadowed the importance of using large Arabic MRC datasets in a pipelined fine-tuning procedure, such as that adopted by top performing systems that also achieved excellent above-median scores [10], [96].

Interestingly, Wasfey, Elrefai, Marwa, *et al.* [136] and Aftab and Malik [9] used part of the Annotated Corpus of Arabic Al-Qur'an Question and Answer (AQQAC) [16] (described in Section 2.3) to augment the *QRCD* training dataset prior to its use in fine-tuning their respective models. They were able to select and exploit about 500-

740 questions from AQQAC; questions were selected only if their respective answers could be extracted from the accompanying verse-based answer. Using the augmented *QRCD* in fine-tuning, the best performing run by Wasfey, Elrefai, Marwa, *et al.* [136] achieved a pRR score well above the median, while the runs by Aftab and Malik [9] attained lower pRR scores well *below* the median. The relatively low performance of [9] could be mainly attributed to the use of the augmented *QRCD* (alone) in fine-tuning an ArabicBERT model [116], as opposed to an AraBERT model that was fine-tuned using additional MSA MRC datasets prior to using the augmented *QRCD* in fine-tuning [136].

Ensemble Learning Trends. From a machine learning perspective, ensemble learning is regarded as the wisdom of the crowd, where multiple models vote towards a prediction [118]. Four systems/runs [51], [136] employed an ensemble of 2-3 different MSA-only pre-trained Arabic BERT-based models. All four runs achieved above median pRR scores, one of which was among the top performing runs [51]. In contrast, a self-ensemble approach was adopted by Premasiri, Ranasinghe, Zaghouani, *et al.* [108] to address the limitation of transformer models being prone to random seed initialization that may cause prediction fluctuations. As such, they trained their models using different random seeds and ensembled the prediction results over those models to ensure more stable predictions.

The above overview has emphasized the relatively good performance of AraELECTRA and AraBERT on our MRC task, which can be further enhanced with an ensemble of both models. It remains to be seen in our future work, how further pre-training AraELECTRA and AraBERTv0.2 using CA resources (with and without ensemble learning) would compare to our CL-AraBERT model.

CHAPTER 3: BUILDING QUR'ANIC QA DATASETS WITH AN EVALUATION PERSPECTIVE

With question answering and machine reading comprehension being an active area of research that intersects with several fields including natural language processing, machine learning, information retrieval, and artificial intelligence, data and language resources for training and testing QA systems are integral and indispensable. The scarcity of Arabic language resources (in comparison to English resources, for example) is the status quo, with the majority of prevalent Arabic resources being in *Modern Standard* Arabic (MSA). In contrary, classical Arabic (CA) resources received little attention, especially for QA and MRC systems on the Holy Qur'an.

Performance evaluation of prevalent Arabic question answering research on the Holy Qur'an was essentially an ad-hoc effort that relied mainly on direct manual judgement of answers returned by the QA system in response to different sets of questions [5], [56], [58], [59], [61], [123], in the absence of having a full set of possible answers to those questions. That precluded the reusability of the judgements for evaluating other QA systems. Rigorous performance comparisons of these QA systems requires publicly available gold standard *test collections*. A test collection is typically composed of a document collection (the Holy Qur'an in our case), a set of queries (questions), and their relevance judgments [80], [132]; the latter is typically a gold standard list of documents that are relevant to each query, as decided through human judgment. For QA test collections, this list constitutes gold standard answers (or verses that have those answers) [132].

To our knowledge, there are no publicly available Arabic question answering test collections on the Holy Qur'an that are fully reusable. For a QA test collection to be *reusable*, it must incorporate a non-trivial coverage of relevant answers to the respective questions [80]. Optimally, when building a QA test collection for the Holy Qur'an (or any religious book for that matter), it should be *fully-reusable* by aiming to include *all* relevant answers to each question.

To address this gap, we introduce *two* datasets that can serve as test collections as well as training CA resources. The first is *AyaTEC*, which is a QA dataset whose answers are qur'anic verses (i.e., verse-based); and the second is *QRCD*, whose answers are spans of text extracted from the accompanying Qur'anic passages that comprise the *direct* verse-based answers of *AyaTEC*. *QRCD* has evolved as an extension of *AyaTEC* to become the first extractive Qur'anic Reading Comprehension Dataset. To the best of effort, the answers to the questions (each represented as a sequence of verses) in *AyaTEC* are exhaustive; i.e., all the qur'anic verses that directly answer the questions were exhaustively extracted and annotated. As such, each of the two datasets fosters a common experimental test-bed for systems to showcase and fairly benchmark their performance.

To facilitate the use of *AyaTEC* and *QRCD* in evaluating Arabic QA and MRC systems on the Holy Qur'an, we propose several evaluation measures to support the different types of questions and the nature of verse-based and span-based answers, while integrating the concept of partial matching of answers in the evaluation.

This chapter is composed of two main parts; the first part is dedicated to the *AyaTEC* dataset, while the second part is dedicated to the *QRCD* dataset. The sections in the first part cover the methodology and the design objectives we adopted in building *AyaTEC*, which is followed by a section to showcase the profile of *AyaTEC* with respect to size, distribution of question types and inter-rater agreement. We conclude the first part by proposing evaluation measures that suit the verse-based nature of the answers in *AyaTEC*. The sections in the second part cover the methodology we adopted in extending *AyaTEC* to develop *QRCD* as a machine reading comprehension dataset. We also conclude the second part by proposing evaluation measures for use with the extractive span-based nature of the answers in *QRCD*.

3.1. Building *AyaTEC*: a Verse-based Test Collection for Arabic QA on the Holy Qur'an

AyaTEC includes 207 questions (with their corresponding 1,762 answers) covering 11 topic categories of the Holy Qur'an that target the information needs of both curious and skeptical users.

Among the main objectives of this work is to build a test collection for the task of evaluating Arabic question answering systems on the Holy Qur'an. Several design objectives were set forth to build *AyaTEC* with the following characteristics.

1. Targeting the information needs of both curious and skeptical users (Section 3.1.1.1). Curious users are defined as those seeking answers to their questions from the Holy Qur'an out of interest in its teachings; and skeptical users as those seeking answers from the Holy Qur'an to questions that may include controversial or undermining issues.

2. Diverse in its topic categories, covering 11 topic categories of the Holy Qur'an (Section 3.1.1.2).

3. Covering factoid and non-factoid questions that are classified into three abstract question types, namely, *single-answer*, *multi-answer* and *no-answer* questions types (described in Section 3.1.1.3).

4. Verse-based, providing answers in the form of qur'anic verses rather than natural language text. Each answer could be a single verse or a set of consecutive verses (Section 3.1.2).

5. Fully reusable. To the best of effort, the set of annotated *direct* answers to each question in the collection was meant to be exhaustive (Section 3.1.2).

6. Large enough to be used in testing and training of systems/models (Section 3.1.4).

In the sub-sections below, we elaborate on how the above design objectives were met.

Our methodology in building *AyaTEC* has followed the typical pipeline of developing test collections, starting with the phase of collecting/developing the questions (as the topics), followed by the relevance judgment phase, as explained in the following sub-sections. We adopted the publicly-available and verified digital version of the Holy Qur'an by Tanzil Project[1] as the source of our document collection of qur'anic verses.

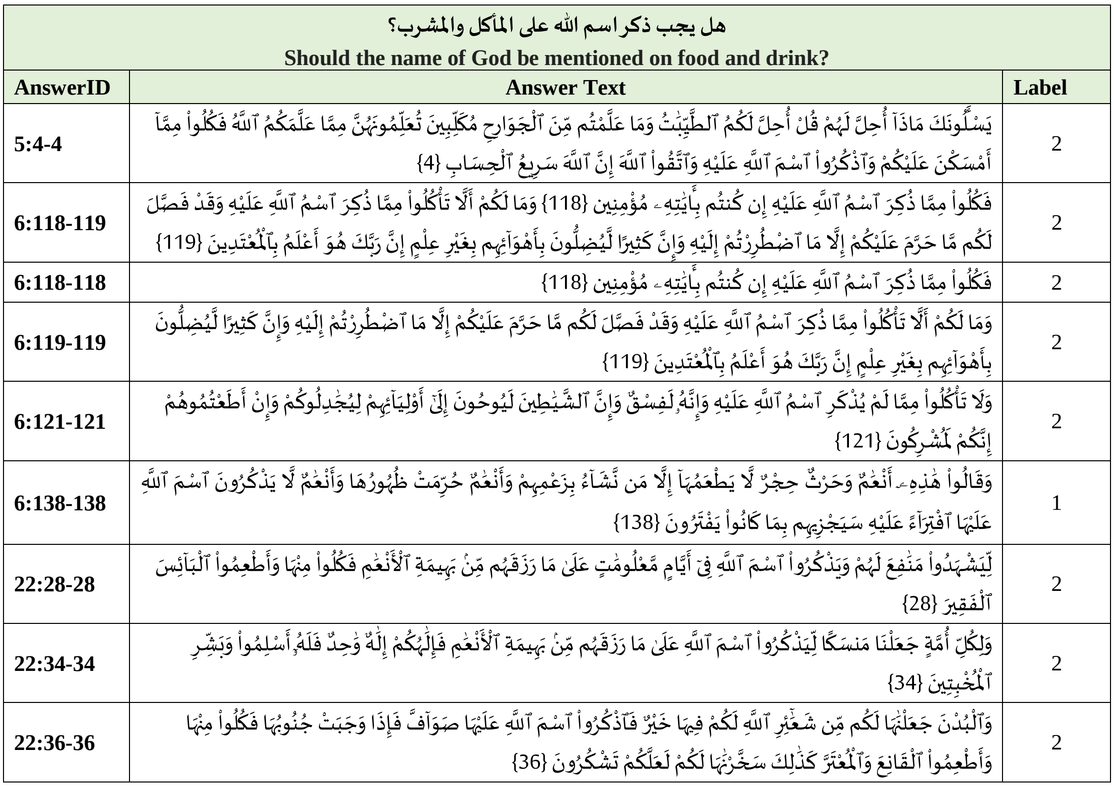
3.1.1. Question Development

Several dimensions were considered while developing the questions set, which include: types of information needs of the target user segments, the topic categories of the Holy Qur'an, and the question types. All questions are assumed to be in MSA (modern standard Arabic), even if a question contains a quoted verse (or partial verse) from the Holy Qur'an.

3.1.1.1. Types of Information Needs

We targeted the information needs of the following two user segments:

[1] http://tanzil.net/docs/Tanzil_Project



	هل يجب ذكر اسم الله على المأكل والمشرب؟ **Should the name of God be mentioned on food and drink?**	
AnswerID	**Answer Text**	**Label**
5:4-4	يَسْـَٔلُونَكَ مَاذَآ أُحِلَّ لَهُمْ قُلْ أُحِلَّ لَكُمُ ٱلطَّيِّبَٰتُ وَمَا عَلَّمْتُم مِّنَ ٱلْجَوَارِحِ مُكَلِّبِينَ تُعَلِّمُونَهُنَّ مِمَّا عَلَّمَكُمُ ٱللَّهُ فَكُلُوا۟ مِمَّآ أَمْسَكْنَ عَلَيْكُمْ وَٱذْكُرُوا۟ ٱسْمَ ٱللَّهِ عَلَيْهِ وَٱتَّقُوا۟ ٱللَّهَ إِنَّ ٱللَّهَ سَرِيعُ ٱلْحِسَابِ {4}	2
6:118-119	فَكُلُوا۟ مِمَّا ذُكِرَ ٱسْمُ ٱللَّهِ عَلَيْهِ إِن كُنتُم بِـَٔايَٰتِهِۦ مُؤْمِنِينَ {118} وَمَا لَكُمْ أَلَّا تَأْكُلُوا۟ مِمَّا ذُكِرَ ٱسْمُ ٱللَّهِ عَلَيْهِ وَقَدْ فَصَّلَ لَكُم مَّا حَرَّمَ عَلَيْكُمْ إِلَّا مَا ٱضْطُرِرْتُمْ إِلَيْهِ وَإِنَّ كَثِيرًا لَّيُضِلُّونَ بِأَهْوَآئِهِم بِغَيْرِ عِلْمٍ إِنَّ رَبَّكَ هُوَ أَعْلَمُ بِٱلْمُعْتَدِينَ {119}	2
6:118-118	فَكُلُوا۟ مِمَّا ذُكِرَ ٱسْمُ ٱللَّهِ عَلَيْهِ إِن كُنتُم بِـَٔايَٰتِهِۦ مُؤْمِنِينَ {118}	2
6:119-119	وَمَا لَكُمْ أَلَّا تَأْكُلُوا۟ مِمَّا ذُكِرَ ٱسْمُ ٱللَّهِ عَلَيْهِ وَقَدْ فَصَّلَ لَكُم مَّا حَرَّمَ عَلَيْكُمْ إِلَّا مَا ٱضْطُرِرْتُمْ إِلَيْهِ وَإِنَّ كَثِيرًا لَّيُضِلُّونَ بِأَهْوَآئِهِم بِغَيْرِ عِلْمٍ إِنَّ رَبَّكَ هُوَ أَعْلَمُ بِٱلْمُعْتَدِينَ {119}	2
6:121-121	وَلَا تَأْكُلُوا۟ مِمَّا لَمْ يُذْكَرِ ٱسْمُ ٱللَّهِ عَلَيْهِ وَإِنَّهُۥ لَفِسْقٌ وَإِنَّ ٱلشَّيَٰطِينَ لَيُوحُونَ إِلَىٰٓ أَوْلِيَآئِهِمْ لِيُجَٰدِلُوكُمْ وَإِنْ أَطَعْتُمُوهُمْ إِنَّكُمْ لَمُشْرِكُونَ {121}	2
6:138-138	وَقَالُوا۟ هَٰذِهِۦٓ أَنْعَٰمٌ وَحَرْثٌ حِجْرٌ لَّا يَطْعَمُهَآ إِلَّا مَن نَّشَآءُ بِزَعْمِهِمْ وَأَنْعَٰمٌ حُرِّمَتْ ظُهُورُهَا وَأَنْعَٰمٌ لَّا يَذْكُرُونَ ٱسْمَ ٱللَّهِ عَلَيْهَا ٱفْتِرَآءً عَلَيْهِ سَيَجْزِيهِم بِمَا كَانُوا۟ يَفْتَرُونَ {138}	1
22:28-28	لِّيَشْهَدُوا۟ مَنَٰفِعَ لَهُمْ وَيَذْكُرُوا۟ ٱسْمَ ٱللَّهِ فِىٓ أَيَّامٍ مَّعْلُومَٰتٍ عَلَىٰ مَا رَزَقَهُم مِّنۢ بَهِيمَةِ ٱلْأَنْعَٰمِ فَكُلُوا۟ مِنْهَا وَأَطْعِمُوا۟ ٱلْبَآئِسَ ٱلْفَقِيرَ {28}	2
22:34-34	وَلِكُلِّ أُمَّةٍ جَعَلْنَا مَنسَكًا لِّيَذْكُرُوا۟ ٱسْمَ ٱللَّهِ عَلَىٰ مَا رَزَقَهُم مِّنۢ بَهِيمَةِ ٱلْأَنْعَٰمِ فَإِلَٰهُكُمْ إِلَٰهٌ وَٰحِدٌ فَلَهُۥٓ أَسْلِمُوا۟ وَبَشِّرِ ٱلْمُخْبِتِينَ {34}	2
22:36-36	وَٱلْبُدْنَ جَعَلْنَٰهَا لَكُم مِّن شَعَٰٓئِرِ ٱللَّهِ لَكُمْ فِيهَا خَيْرٌ فَٱذْكُرُوا۟ ٱسْمَ ٱللَّهِ عَلَيْهَا صَوَآفَّ فَإِذَا وَجَبَتْ جُنُوبُهَا فَكُلُوا۟ مِنْهَا وَأَطْعِمُوا۟ ٱلْقَانِعَ وَٱلْمُعْتَرَّ كَذَٰلِكَ سَخَّرْنَٰهَا لَكُمْ لَعَلَّكُمْ تَشْكُرُونَ {36}	2

Figure 3.1. A *single-answer* question raised by a curious user and its exhaustive set of direct answers in the Holy Qur'an. AnswerID has the form *Chapter#:StartVerse#-EndVerse#.* Labels 2 and 1 correspond to direct and indirect answers, respectively.

1. ***Curious users*** seeking answers from the Holy Qur'an to their questions, out of interest in its teachings.

2. ***Skeptical users*** seeking answers from the Holy Qur'an to questions that may include controversial or undermining questions.

To cater for the first type of information needs (for curious users), we acquired a total of 145 Arabic questions; 99 out of them were used in evaluating two Arabic question answering systems on the Holy Qur'an: 54 from Abdelnasser, Ragab, Mohamed, *et al.* [5] and 45 from Hakkoum and Raghay [56]. The remaining 46 (out of the 145) questions were acquired by soliciting questions from users directly. Questions with a huge answer space were excluded because they would incur a very high annotation cost, given that

من هم الأنبياء الذين ذكروا في القرآن على أنهم مسلمون؟ Who are the prophets that were mentioned in the Quran as being Muslims?		
AnswerID	**Answer Text**	**Label**

2:127-128	وَإِذْ يَرْفَعُ إِبْرَٰهِيمُ ٱلْقَوَاعِدَ مِنَ ٱلْبَيْتِ وَإِسْمَٰعِيلُ رَبَّنَا تَقَبَّلْ مِنَّآ إِنَّكَ أَنتَ ٱلسَّمِيعُ ٱلْعَلِيمُ {127}
	رَبَّنَا وَٱجْعَلْنَا مُسْلِمَيْنِ لَكَ وَمِن ذُرِّيَّتِنَآ أُمَّةً مُّسْلِمَةً لَّكَ وَأَرِنَا مَنَاسِكَنَا وَتُبْ عَلَيْنَآ إِنَّكَ أَنتَ ٱلتَّوَّابُ ٱلرَّحِيمُ {128}
		2

2:132-132	وَوَصَّىٰ بِهَآ إِبْرَٰهِيمُ بَنِيهِ وَيَعْقُوبُ يَٰبَنِيَّ إِنَّ ٱللَّهَ ٱصْطَفَىٰ لَكُمُ ٱلدِّينَ فَلَا تَمُوتُنَّ إِلَّا وَأَنتُم مُّسْلِمُونَ {132}	2
2:133-133	أَمْ كُنتُمْ شُهَدَآءَ إِذْ حَضَرَ يَعْقُوبَ ٱلْمَوْتُ إِذْ قَالَ لِبَنِيهِ مَا تَعْبُدُونَ مِنۢ بَعْدِى قَالُوا۟ نَعْبُدُ إِلَٰهَكَ وَإِلَٰهَ ءَابَآئِكَ إِبْرَٰهِيمَ وَإِسْمَٰعِيلَ وَإِسْحَٰقَ إِلَٰهًا وَٰحِدًا وَنَحْنُ لَهُۥ مُسْلِمُونَ {133}	2
2:136-136	قُولُوٓا۟ ءَامَنَّا بِٱللَّهِ وَمَآ أُنزِلَ إِلَيْنَا وَمَآ أُنزِلَ إِلَىٰٓ إِبْرَٰهِيمَ وَإِسْمَٰعِيلَ وَإِسْحَٰقَ وَيَعْقُوبَ وَٱلْأَسْبَاطِ وَمَآ أُوتِىَ مُوسَىٰ وَعِيسَىٰ وَمَآ أُوتِىَ ٱلنَّبِيُّونَ مِن رَّبِّهِمْ لَا نُفَرِّقُ بَيْنَ أَحَدٍ مِّنْهُمْ وَنَحْنُ لَهُۥ مُسْلِمُونَ {136}	2
3:67-67	مَا كَانَ إِبْرَٰهِيمُ يَهُودِيًّا وَلَا نَصْرَانِيًّا وَلَٰكِن كَانَ حَنِيفًا مُّسْلِمًا وَمَا كَانَ مِنَ ٱلْمُشْرِكِينَ {67}	2
3:84-84	قُلْ ءَامَنَّا بِٱللَّهِ وَمَآ أُنزِلَ عَلَيْنَا وَمَآ أُنزِلَ عَلَىٰٓ إِبْرَٰهِيمَ وَإِسْمَٰعِيلَ وَإِسْحَٰقَ وَيَعْقُوبَ وَٱلْأَسْبَاطِ وَمَآ أُوتِىَ مُوسَىٰ وَعِيسَىٰ وَٱلنَّبِيُّونَ مِن رَّبِّهِمْ لَا نُفَرِّقُ بَيْنَ أَحَدٍ مِّنْهُمْ وَنَحْنُ لَهُۥ مُسْلِمُونَ {84}	2

6:161-163	قُلْ إِنَّنِى هَدَىٰنِى رَبِّىٓ إِلَىٰ صِرَٰطٍ مُّسْتَقِيمٍ دِينًا قِيَمًا مِّلَّةَ إِبْرَٰهِيمَ حَنِيفًا وَمَا كَانَ مِنَ ٱلْمُشْرِكِينَ {161}
	قُلْ إِنَّ صَلَاتِى وَنُسُكِى وَمَحْيَاىَ وَمَمَاتِى لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ {162} لَا شَرِيكَ لَهُۥ وَبِذَٰلِكَ أُمِرْتُ وَأَنَا۠ أَوَّلُ ٱلْمُسْلِمِينَ {163}
		1

10:71-72	وَٱتْلُ عَلَيْهِمْ نَبَأَ نُوحٍ إِذْ قَالَ لِقَوْمِهِۦ يَٰقَوْمِ إِن كَانَ كَبُرَ عَلَيْكُم مَّقَامِى وَتَذْكِيرِى بِـَٔايَٰتِ ٱللَّهِ فَعَلَى ٱللَّهِ تَوَكَّلْتُ فَأَجْمِعُوٓا۟ أَمْرَكُمْ وَشُرَكَآءَكُمْ ثُمَّ لَا يَكُنْ أَمْرُكُمْ عَلَيْكُمْ غُمَّةً ثُمَّ ٱقْضُوٓا۟ إِلَىَّ وَلَا تُنظِرُونِ {71} فَإِن تَوَلَّيْتُمْ فَمَا سَأَلْتُكُم مِّنْ أَجْرٍ إِنْ أَجْرِىَ إِلَّا عَلَى ٱللَّهِ وَأُمِرْتُ أَنْ أَكُونَ مِنَ ٱلْمُسْلِمِينَ {72}	2
10:71-71	وَٱتْلُ عَلَيْهِمْ نَبَأَ نُوحٍ إِذْ قَالَ لِقَوْمِهِۦ يَٰقَوْمِ إِن كَانَ كَبُرَ عَلَيْكُم مَّقَامِى وَتَذْكِيرِى بِـَٔايَٰتِ ٱللَّهِ فَعَلَى ٱللَّهِ تَوَكَّلْتُ فَأَجْمِعُوٓا۟ أَمْرَكُمْ وَشُرَكَآءَكُمْ ثُمَّ لَا يَكُنْ أَمْرُكُمْ عَلَيْكُمْ غُمَّةً ثُمَّ ٱقْضُوٓا۟ إِلَىَّ وَلَا تُنظِرُونِ {71}	1
12:99-101	فَلَمَّا دَخَلُوا۟ عَلَىٰ يُوسُفَ ءَاوَىٰٓ إِلَيْهِ أَبَوَيْهِ وَقَالَ ٱدْخُلُوا۟ مِصْرَ إِن شَآءَ ٱللَّهُ ءَامِنِينَ {99} وَرَفَعَ أَبَوَيْهِ عَلَى ٱلْعَرْشِ وَخَرُّوا۟ لَهُۥ سُجَّدًا وَقَالَ يَٰٓأَبَتِ هَٰذَا تَأْوِيلُ رُءْيَٰىَ مِن قَبْلُ قَدْ جَعَلَهَا رَبِّى حَقًّا وَقَدْ أَحْسَنَ بِىٓ إِذْ أَخْرَجَنِى مِنَ ٱلسِّجْنِ وَجَآءَ بِكُم مِّنَ ٱلْبَدْوِ مِنۢ بَعْدِ أَن نَّزَغَ ٱلشَّيْطَٰنُ بَيْنِى وَبَيْنَ إِخْوَتِىٓ إِنَّ رَبِّى لَطِيفٌ لِّمَا يَشَآءُ إِنَّهُۥ هُوَ ٱلْعَلِيمُ ٱلْحَكِيمُ {100} رَبِّ قَدْ ءَاتَيْتَنِى مِنَ ٱلْمُلْكِ وَعَلَّمْتَنِى مِن تَأْوِيلِ ٱلْأَحَادِيثِ فَاطِرَ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ أَنتَ وَلِىِّۦ فِى ٱلدُّنْيَا وَٱلْءَاخِرَةِ تَوَفَّنِى مُسْلِمًا وَأَلْحِقْنِى بِٱلصَّٰلِحِينَ {101}	2
27:41-44	قَالَ نَكِّرُوا۟ لَهَا عَرْشَهَا نَنظُرْ أَتَهْتَدِىٓ أَمْ تَكُونُ مِنَ ٱلَّذِينَ لَا يَهْتَدُونَ {41} فَلَمَّا جَآءَتْ قِيلَ أَهَٰكَذَا عَرْشُكِ قَالَتْ كَأَنَّهُۥ هُوَ وَأُوتِينَا ٱلْعِلْمَ مِن قَبْلِهَا وَكُنَّا مُسْلِمِينَ {42} وَصَدَّهَا مَا كَانَت تَّعْبُدُ مِن دُونِ ٱللَّهِ إِنَّهَا كَانَتْ مِن قَوْمٍ كَٰفِرِينَ {43} قِيلَ لَهَا ٱدْخُلِى ٱلصَّرْحَ فَلَمَّا رَأَتْهُ حَسِبَتْهُ لُجَّةً وَكَشَفَتْ عَن سَاقَيْهَا قَالَ إِنَّهُۥ صَرْحٌ مُّمَرَّدٌ مِّن قَوَارِيرَ قَالَتْ رَبِّ إِنِّى ظَلَمْتُ نَفْسِى وَأَسْلَمْتُ مَعَ سُلَيْمَٰنَ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ {44}	2

Figure 3.2. A *multi-answer* question raised by a skeptical user; only a sample of the answers are shown. Labels 2 and 1 correspond to direct and indirect answers, respectively.

AyaTEC should exhaustively include all the potential answers (qur'anic verses) to each considered question. Figure 3.1 shows a sample question from this segment.

To address the second type of information needs (for skeptical users), we acquired 62 Arabic questions using two methods: soliciting questions from users directly, and drawing questions from YouTube videos and books. Unfortunately, a large number of the questions collected initially from skeptical users did not have an answer from the Holy Qur'an. As such, some of these questions were slightly rephrased, others were deleted, and a fraction not exceeding 15% of the total number of questions in *AyaTEC* were purposely retained to add a flair of challenge to the test collection, similar to the English dataset developed by Rajpurkar, Jia, and Liang [110]. Figure 3.2 illustrates a sample question from this category.

3.1.1.2. Topic Categories of the Holy Qur'an

To have a wide topical coverage, we chose our questions from 11 different topical categories of the Holy Qur'an. The developed questions covered these topic categories in different proportions, with the biggest share of questions being on *Provisions of Islam* and *Stories of Prophets*, followed by *Former Nations*, as shown in Figure 3.3.

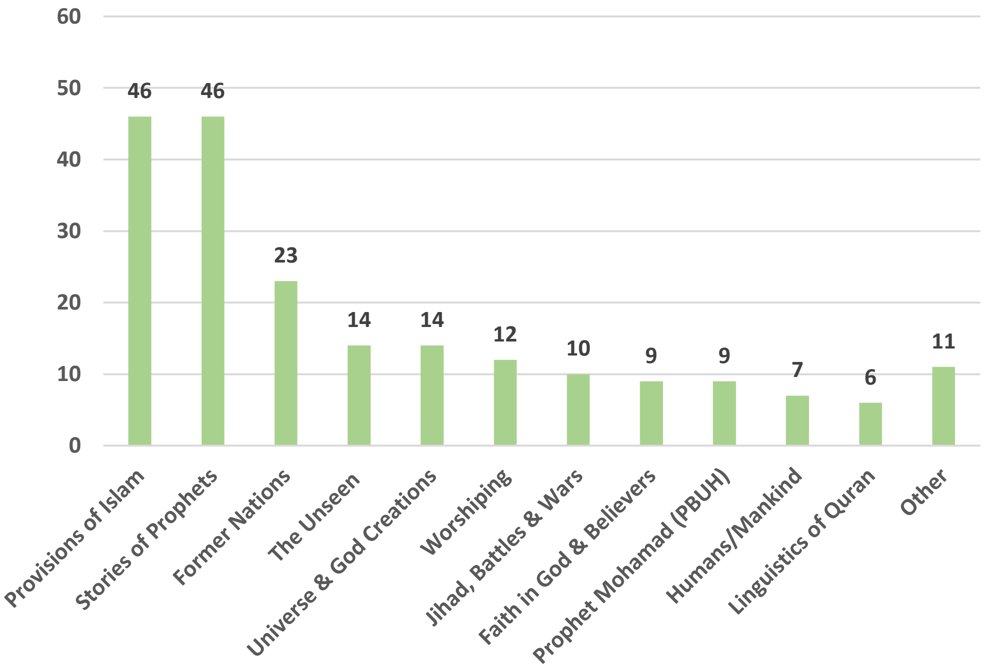


Figure 3.3. Distribution of questions over 11 Holy Qur'an topic categories.

3.1.1.3. Question Types

Although the questions include a variety of factoid, list, definition, causal, relation, and yes/no questions, we have adopted the abstracted classification of *single-answer* and *multi-answer* questions, which correspond to having one answer or multiple answers in the Qur'an, respectively. This is due to the nature of the answers being qur'anic verses rather than traditional natural language answers. For example, when answering a yes/no question from the Holy Qur'an, the answer should include all the verses that would provide distinct evidence that supports a *Yes* answer or a *No* answer. If the answer to the given yes/no question constitutes more than one evidence, that would make it a multi-answer question, otherwise, it will be a single-answer question. This may also apply on the other question types (factoid, casual, etc.). As such, the single/multi-answer classification was adopted to encompass any question type. Additionally, a *no-answer* question type was defined to cater for the questions that do *not* have an answer in the Holy Qur'an. This classification has indeed an implication on the evaluation schemes to adopt, as discussed in Section 3.1.6. The question types are formally defined as follows:

- ***Single-Answer*** questions are those having only *one* answer in the Holy Qur'an. The answer (qur'anic verse/verses) may be repeated (in different positions) in the Holy Qur'an. Naturally, the repeated answers could be syntactically and/or semantically similar. Figure 3.1 exhibits an example of this question type.

- ***Multi-Answer*** questions are those having two or more *different* answers, or those with an answer that constitutes several components. Each answer may also be repeated in the Holy Qur'an. Figure 3.2 exhibits an example for this question type.

- ***No-Answer*** questions are of two types; namely, *zero-answer* questions and *no-direct-answer* questions. *Zero-answer* questions are those that have *no* answer in the Holy Qur'an, while *no-direct-answer* questions are those that do not have a *direct* answer (i.e., questions that only have *indirect* answers). An answer is *direct* if it responds to a given question *explicitly*, and its *context* is *consistent* with the context of the question; otherwise, the answer is *indirect*. Formal definitions of *direct* and *indirect* answers are provided in Section 3.1.2.2 with examples.

3.1.2. Relevance Judgements

Given the sensitivity of dealing with a sacred book, and our aim to include, in *AyaTEC*, exhaustive sets of all the answer occurrences of qur'anic verses that would potentially answer every question, three specialists in Holy Qur'an Interpretation (Tafseer) were sought for the relevance judgments phase; each was expected to have completely memorized the Qur'an.

Initially, we conducted a pilot experiment on 10 questions drawn from the developed question set, such that they include samples from the different question types. One of the three specialists was requested to answer these questions by extracting all the potential verses that answer every question and then annotate each answer based on a given rubric as *direct*, *indirect*, or *incorrect* as elaborated in Section 3.1.2.2. The task was overwhelmingly very time consuming and difficult. To make the task easier, we decided to partition the task into two separate steps: *answer extraction* and *answer annotation*.

3.1.2.1. Answer Extraction

Two UpWork[2] freelancers who are very knowledgeable in Qur'an were hired to extract all the potential relevant answers (qur'anic verses) to a given question. They were advised to use the search tools of KSU's Electronic Moshaf Project[3] and/or the Tanzil Project.[4] Their competence was verified before the hire using the 10 pilot questions described above. The freelancer was hired if 90% of the gold-standard set of previously extracted answers for the pilot questions were achieved. The question set was eventually distributed evenly among the two answer extractors.

The answer extraction step followed the following guidelines:

- One occurrence of an answer could be a single verse or a set of consecutive verses (2 to 10 verses).
- To the best of effort, all the repeated occurrences of the answer in the Holy Qur'an must be extracted.
- If the answer extractors face a challenging question, they may leave it to the Qur'an specialists to extract its answer(s), if they exist.
- The answer to a multi-answer question can belong to one of the following cases.
  - One single verse contains all the constituents of the answer. We refer to each constituent as an answer *instance*. Figure 3.4 exhibits examples of answer *instances*.
  - Several single verses; each may contain one or more instances of the answer.

[2] https://www.upwork.com/

[3] http://quran.ksu.edu.sa/

[4] http://tanzil.net/

  - A set of consecutive verses (2 to 10 verses) may contain one or more instances of the answer. This implies that some verses are part of the answer despite the fact that they may not include any instances of the answer. Such verses are retained to elaborate the context of the answer.
  - A combination of single verses and sets of consecutive verses that may contain one or more instances of the answer; i.e., a combination of the previous two cases above.

- A single verse that is a constituent of an answer will have a *verse-ID* of the form *Chapter#:Verse#*. For example, the verse-IDs constituting the second answer in Figure 3.1 are *6:118* and *6:119*, respectively.

- Consequently, each answer in *AyaTEC* has an *answer-ID* that is of the form *Chapter#:StartVerse#-EndVerse#*. For example, the first two answers in Figure 3.1 have the answer-IDs *5:4-4* and *6:118-119*, respectively.

3.1.2.2. Answer Annotation

With the answer extraction task completed, each of the three Qur'an specialists evaluate and annotate *all* the extracted answers of *all* the questions (to facilitate majority voting), taking into consideration the following guidelines:

- The extracted answers must be checked to verify that all the answers to a given question and their occurrences in the Holy Qur'an have been extracted; if not, missing answers must be added.

- Each extracted answer must be evaluated and labeled as *Direct*, *Indirect*, or *Incorrect* as defined below:

  - **Direct**: If the extracted answer responds to the given question *explicitly*, and the *context* of the answer is *consistent* with the context of the question. Examples of direct answers are those with a label/annotation code of '2' in Figures 3.1 and 3.2.
  - **Indirect**: If the extracted answer complies with one of the following cases: (a) it answers the given question explicitly, *but* the context of the answer is *inconsistent* with the context of the question; or (b) it answers the given question *implicitly* and the context of the answer is consistent with the context of the question. Examples of indirect answers are those with a label/annotation code of '1' in Figures 3.1 and 3.2.
  - **Incorrect**: If the extracted answer does not answer the question.

- In the case where the annotator encounters a question that does not have any answer from the Holy Qur'an, it would be designated as a *zero-answer* question.

- As each Qur'an specialist may discover different answers or answer occurrences that were not extracted by the UpWork answer extractors, it was imperative to synchronize the newly discovered answers across the three Qur'an specialists. Each specialist must evaluate all the newly discovered answer occurrences so as to apply majority voting.

While *AyaTEC* was developed to include an exhaustive set of *all direct* answers per question, the *indirect* answers that are included may not be exhaustive. This is due to their propensity to be highly ubiquitous and intractable for some questions, given the high presence of anaphoric-structures in the Holy Qur'an. Nevertheless, we retained the annotated *indirect* answers (even if they are not considered in the evaluation) for two reasons: 1) to differentiate between questions that do not have an answer from the Holy Qur'an (zero-answer type) and those that do not have a direct answer (no-direct-answer type), and 2) to cater for the types of information needs of curious/skeptical users that would presumably be partially satisfied to know whether the Holy Qur'an has, or does not have thereof, direct/indirect answers to their questions. Hence, in the absence of direct answers, the *indirect* answers would act as clues for curious or skeptical users to further explore in renowned Interpretation (Tafseer) books of the Holy Qur'an.

3.1.3. Post-Annotation Processing

We present next the processing steps applied on the annotated answers after completing the answer-annotation task.

3.1.3.1. Majority Voting

Majority voting was applied on the QA pairs whenever there were, at least, two inter-annotator agreements. For cases with no inter-annotator agreement or a missing annotation, a fourth Qur'an specialist was sought to break the tie, or compensate for the missing annotation, respectively.

3.1.3.2. Excluding Redundant Answers

To eliminate the possibility of having redundant answers of a given question that carry the same label, we devised a set of rules that were applied on the annotated QA pairs (after majority voting) to designate each answer (QA pair) with its final label as direct, indirect or incorrect. Ensuring no redundancy between the answers (that carry the same label) has an implication on the fairness of the evaluation measures we propose in Section 3.1.6. Different rules were developed for single-answer and multi-answer questions respectively.

1. **For Single-Answer Questions**:

   We retained the *tightest* annotated answers (i.e., answers having the smallest number of verses) for each single-answer question, and deleted any super-sets of those retained answers that carry the same label; therefore, each retained answer may not include answer subsets or super-sets annotated with the same label.

   For example, applying this rule on the QA pairs in Figure 3.1 results in removing the direct answer-ID *6:118-119* (whose label=2) while retaining the direct answer-IDs *6:118-118* and *6:119-119* that carry a label of '2' as well.

2. **For Multi-Answer Questions**:

   (a) We retained the *widest* annotated answers (i.e., answers having the largest number of verses) for each multi-answer question, and deleted any subsets of those retained answers according to the following policy:

      i. If the *widest* retained answer's final label is *direct*, all answer subsets are deleted regardless of their carried label. This rule is applied to address a downside for applying the annotation guidelines in Section 3.1.2.2, where an answer subset may not be qualified to be a direct answer on its own, although it may contain a correct instance of the answer. For example, applying this rule on the QA pairs in Figure 3.2 results in removing the indirect answer-ID *10:71-71* (whose label=1) while retaining the direct answer-ID *10:71-72*.

      ii. If the *widest* retained answer's final label is *indirect* or *incor-*

rect, only answer subsets carrying the same respective labels are deleted.

   (b) Any verse/answer subset included in a *direct widest* retained answer will implicitly carry a direct label as well.

3. For no-direct-answer questions, we adopted the rules applied on multi-answer-questions.

Applying the above rules reduced the number of annotated question-answer pairs in *AyaTEC* from 2064 to 1762.

3.1.3.3. Development of Answer-Instance Sets and Verse-to-Instances Maps

For the purpose of evaluating multi-answer questions, two additional data components were developed for each multi-answer question: an answer-instance set and a verse-to-instances map. The answer-instance set contains the gold answer instances for a given multi-answer question, while the verse-to-instances map encodes the distribution of the gold answer instances among the verses that constitute the gold direct answers for that question. Details on the use of these two data components in evaluation are presented in Section 3.1.6.3.

Answer-Instance Set. Answer sets of distinct constituents/instances were developed for each question with the help of one of the Qur'an specialists. The specialist extracted the answer-instance sets from the final list of annotated answers (after majority voting), where only direct answers were considered. Figure 3.4 exhibits the answer-instance set for an example question.

من هم الأنبياء الذين ذكروا في القرآن على أنهم مسلمون؟ **Who are the prophets that were mentioned in the Quran as being Muslims?**		
1	Ibrahim/Abraham	إِبْرَٰهِيم
2	Ismail	إِسْمَٰعِيل
3	Yaqub/Jacob	يَعْقُوب
4	Ishaq/Isaac	إِسْحَٰق
5	Descendants	ٱلْأَسْبَاطِ
6	Moussa/Moses	مُوسَىٰ
7	Essa/Jesus	عِيسَى
8	Nooh/Noah	نُوح
9	Yousuf/Joseph	يُوسُف
10	Sulaiman/Solomon	سُلَيْمَٰن

Figure 3.4. An example answer-instance set for a multi-answer question.

Verse-to-Instances Map. This map stores the answer instances indicated by each verse in a direct answer to a multi-answer question. Ideally, for a verse to be included in the map, it should include one or more answer instances. However, we encountered cases where several verses in an answer may contribute to one answer instance. For this reason, the verse-to-instances map may include verses not contributing directly to an instance. Figure 3.5 depicts an example map for the same question above.

من هم الأنبياء الذين ذكروا في القرآن على أنهم مسلمون؟ **Who are the prophets that were mentioned in the Quran as being Muslims?**			
Answer-ID	**Verse-ID**	**No. of Instances**	**Instances**
2:127-128	**2:127**	2	إِبْرَٰهِيم، إِسْمَٰعِيل
2:132-132	**2:132**	2	إِبْرَٰهِيم، يَعْقُوب
2:133-133	**2:133**	4	يَعْقُوب، إِبْرَٰهِيم، إِسْمَٰعِيل، إِسْحَٰق
2:136-136	**2:136**	7	إِبْرَٰهِيم، وإِسْمَٰعِيل، وإِسْحَٰق ، ويَعْقُوب، وٱلْأَسْبَاطِ، ومُوسَىٰ، وعِيسَى
3:67-67	**3:67**	1	إِبْرَٰهِيم
3:84-84	**3:84**	7	إِبْرَٰهِيم، وإِسْمَٰعِيل، وإِسْحَٰق ، ويَعْقُوب، وٱلْأَسْبَاطِ، ومُوسَىٰ، وعِيسَى
10:71-72	**10:71**	1	نُوح
12:99-101	**12:99**	1	يُوسُف
27:41-44	**27:44**	1	سُلَيْمَٰن

Figure 3.5. An example verse-to-instances map for a multi-answer question. Verse-IDs that constitute direct answers but do not contribute to any answer instance are not shown.

3.1.4. Profile of AyaTEC

In this section, we describe the profile of *AyaTEC* quantitatively before presenting an inter-rater agreement analysis.

AyaTEC is composed of 207 questions; 145 questions target the information needs of curious users (70%), while 62 questions target the information needs of skeptical users (30%). We decided to include the latter type of questions because none of the prevalent Arabic QA research on the Holy Qur'an, nor the QA test collections mentioned in Section 2.1, have catered for questions that target the information needs of skeptical users.

The questions cover 11 qur'anic topic categories as shown in Figure 3.3. It was expected for the topic categories of *Provisions of Islam* and *Stories of Prophets* to have relatively larger shares of the questions (22% each) than the other topic categories.

Figure 3.6 depicts the distribution of the questions by question type. Single-answer and multi-answer questions have relatively comparable shares, whereas no-answer questions comprise 17% (15% are zero-answer questions, and only 2% are no-direct-answer questions), as explained in Section 3.1.1.3.

It is worth noting that among the 62 questions that target the information needs of skeptical users, 15 questions (24% of the 62) are of zero-answer type. In contrast, only 16 questions (11%) out of the 145 that target the information needs of curious users are zero-answer questions.

AyaTEC includes the annotations of the three Qur'an specialists over a total of 1,762 question-answer pairs for 176 questions that have answers. As indicated in Section 3.1.3.1, majority voting was applied whenever there were, at least, two inter-annotator agreements; and for tie-breaking or missing annotations, a fourth Qur'an

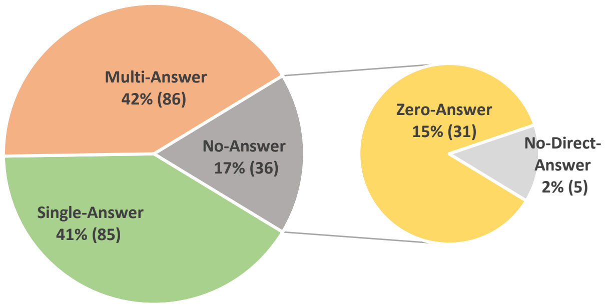


Figure 3.6. Distribution of questions in *AyaTEC* by question type.

specialist was sought.

Figure 3.7 shows the distribution of the resulting majority votes among *direct*, *indirect*, and *incorrect* question-answer pairs (or answers for short). *Direct* answers had the biggest share (67%).

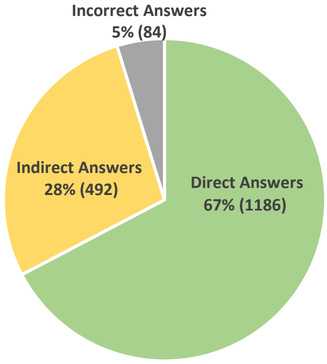


Figure 3.7. Distribution of the annotated question-answer pairs in *AyaTEC* by label type.

Single-answer and multi-answer questions have 534 and 1,204 QA pairs (answers), respectively. Direct answers constitute 42% of the answers for single-answer questions, and a bigger share of 80% of the answers for multi-answer questions. Consequently, the average number of direct answers for single-answer questions and multi-answer questions is 3 and 11, respectively.

With quality being a main concern while building *AyaTEC*, we present next an analysis of the strength of inter-annotator agreement among the three Qur'an specialists.

3.1.5. Inter-Rater Agreement

The Fleiss' kappa coefficient was used to assess inter-annotator agreement among the three Qur'an specialists (Table 3.1). Fleiss' kappa is an extension of the Cohen's Kappa coefficient; the latter is used in the literature to measure agreement between two raters only, where agreement due to chance is factored out. We have chosen Fleiss' kappa because it can measure agreement among more than two raters [124]. The following proposed interpretation of the Kappa statistic by Landis and Koch [76] was adopted to indicate the strength of agreement: $<=$ 0.0 is poor, 0.01-0.20 is slight, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is substantial, and 0.81-1.0 is almost perfect. However, some controversy exists in the literature towards the accurateness of this interpretation, given that the number of rating categories and the number of subjects (QA pairs in our case) can adversely affect the value of kappa [124].[5]

[5] We could not apply Fleiss' Kappa on a small number of QA pairs/subjects, as shown in Table 3.1 for no-direct-answer questions.

Table 3.1. Fleiss Kappa scores for inter-annotator agreement among the three Qur'an specialists over QA pairs.

Question	QA Pairs #	Kappa-All	Kappa-Direct
All	1,762	0.31	0.42
By Question Type			
Single-answer	534	0.37	**0.58**
Multi-answer	1204	0.19	0.23
No-direct-answer	24	-	-
By User Category			
Curious	1066	0.47	**0.61**
Skeptical	696	0.09	0.15
Multi-answer (Curious questions)	681	0.38	0.44

According to the kappa scores exhibited in Table 3.1, and the kappa interpretation scale mentioned above, inter-annotator agreement seems to be fair (0.31) over all the 1,762 QA pairs in the collection, and moderate (0.42) when computed only over the *direct* evaluation ratings. Henceforward, we will refer to the kappa scores computed over *all* ratings (including direct, indirect and incorrect ratings) as kappa-all, whereas we refer to the kappa scores computed over the *direct* ratings only as kappa-direct. We justify our emphasis on kappa-direct scores rather than kappa-all scores for three reasons: (1) the core value of *AyaTEC* is in its use to evaluate QA systems over direct answers, (2) kappa-all scores show a high resemblance to the pattern of behavior of kappa-direct scores across question types and user categories, (3) kappa-direct scores are significantly better than kappa-all scores.

It is worth noting that the Kappa-direct scores were significantly higher when computed over the QA pairs of questions that target the information needs of curious users (0.61), and the QA pairs of single-answer questions (0.58). On the other hand, the kappa-direct scores were significantly lower over the QA pairs of questions that target the information needs of skeptical users (0.15), and the QA pairs of multi-answer questions (0.23). These results suggest that inter-rater *disagreement* is mainly attributed to the questions of skeptical users. This finding was ascertained by the witnessed significant increase of the kappa-direct score (from 0.23 to 0.44) when computed over the multi-answer questions excluding those that target skeptical users (Table 3.1). As such, disagreement among the Qur'an specialists over this kind of questions is not considered surprising, due to the arguable nature of some of these questions and hence answers.

3.1.6. Using AyaTEC *in Evaluation*

In this section, we discuss how one can use *AyaTEC* to evaluate a QA system for the Holy Qur'an that returns answers in terms of verses. We assume a ranked retrieval setting where the system returns a ranked list of answers to a given question; each answer is a sequence of one or more consecutive verses. The evaluation could be designed to support two user satisfaction scenarios. In the first scenario, the user would be satisfied to get *any one* occurrence of an answer to his/her question from the system; as such, the repeated occurrences of the answer can be ignored in the evaluation. In the second scenario, the user would anticipate getting *all* occurrences of an answer to his/her question. For both scenarios, we focus our evaluation on *direct* answers exclusively, assuming the user is pursuing *only* direct answers whenever they exist, since it is infeasible to track all potential indirect answers in the Holy Qur'an.

We propose possible evaluation measures to adopt for each of the three types of questions, namely, single-answer, multi-answer and no-answer questions. For all proposed measures, we will discuss the evaluation of system answers given one question. The overall evaluation score is indeed the average over all questions of the corresponding type in *AyaTEC*. We assume that the system retrieves a ranked list R of answers, which is evaluated against the set of gold direct answers A for the given question.

3.1.6.1. Partial Matching of Answers

Recall that each answer (either gold or returned by the system) may constitute one or more consecutive verses (up to 10). To give credit to QA systems that may retrieve an answer that does not fully match one of the gold answers, but partially matches it, we introduce the notion of *partial matching* of answers. We define the answer-matching score m of a system answer r, denoted by m_r, as the maximum matching score of r over all direct gold answers A of the question.

$$m_r = \max_{a \in A} f_m(r, a) \tag{3.1}$$

where $f_m(r, a)$ is an answer matching function that matches a system answer r with a correct direct answer a. Inspired by Rajpurkar, Jia, and Liang [110] and Rajpurkar, Zhang, Lopyrev, *et al.* [111], we propose using the F_1 measure, applied over verse-IDs, as the answer-matching function. F_1 score is the harmonic mean of precision and recall applied over the verse-IDs that constitute answers (rather than tokens constituting textual answers as in [110], [111]). In this case, we treat the answers as bags of verse-IDs and compute the precision and recall (and hence F_1) accordingly.

$$f_m(r, a) = F_1(V_r|V_a) \tag{3.2}$$

where V_r and V_a are the sequences (treated as sets) of verses constituting the system and gold answers, respectively.[6] Refer to Figure A.1 in Appendix A for an example of partial matching computation.

3.1.6.2. Evaluating Single-Answer Questions

Having defined the answer matching score, we present next the proposed evaluation measures to use in evaluating the system answers given a single-answer question.

3.1.6.2.1. Retrieving any occurrence of the answer. The first measure is a variant of *Precision@1* measure, but considering partial matches. We denote it by *Matching@1*, or shortly $M@1$.

$$M@1(R) = m_{r_1} \tag{3.3}$$

where m_{r_1} is the matching score of the answer at the first rank. This measure only looks at the first returned answer.

The second measure is a variant of the *Reciprocal Rank* (RR) measure, but considering partial matches. We denote it by *Partial Reciprocal Rank*, or shortly pRR.

$$pRR(R) = \frac{m_{r_k}}{k} \; ; \; k = \min\{k \mid m_{r_k} > 0\} \tag{3.4}$$

where k refers to the rank position of the *first* answer that has non-zero matching score.

[6] Alternatively, any measure of set overlap (e.g., Jaccard Coefficient) can also be used.

This one gives partial credit for systems that return a relevant answer not necessarily at the top of the ranked list. Alternatively, one can also limit k to be the rank of the first answer with a matching score of 1, i.e., an exact-match.

Note that, in both measures, the matching score is used in lieu of a binary score in the corresponding traditional measures, *Precision@1* and *RR*, respectively.

3.1.6.2.2. Retrieving all occurrences of the answer. In the same spirit of the previous measures, we propose using variants of the *Recall* and the *Precision* measures that consider partial matches to evaluate the system answers for the task of retrieving all occurrences of the single answer. We denote them by *Partial Recall* ($pRecall$) and *Partial Precision* ($pPrecision$), respectively.

$$pRecall(R) = \frac{\sum_{r \in R} m_r}{|A|} \tag{3.5}$$

$$pPrecision(R) = \frac{\sum_{r \in R} m_r}{|R|} \tag{3.6}$$

where m_r is the answer-matching score of a system answer r as defined in equation 3.1, and $|A|$ and $|R|$ are the sizes of the gold and returned answers A and R, respectively.

We note that in computing the answer matching scores, returned system answers are matched in the order of their ranking, such that no gold direct answer a in the set A is best-matched to more than one system answer r. We then compute F_1 as the performance measure of the system given the question, applied over $pRecall$ and $pPrecision$.

3.1.6.3. Evaluating Multi-Answer Questions

Multi-answer questions in *AyaTEC* include a variety of list, definition, causal, and relation questions. Each question has its corresponding answer-instance set (see Figure 3.4 for an example). For this type of questions, the system is required to return answers that cover *all* the answer-instances for a given question. For the first satisfaction scenario, redundant occurrences of the same answer instance are ignored. On the contrary, the second satisfaction scenario requires the retrieval of all occurrences of each answer instance.

3.1.6.3.1. Retrieving any occurrence of answer instances. In this scenario, the system returns a ranked list of answers aiming to cover all answer instances by retrieving at least one occurrence of each, while retrieving a minimal set of answers to maintain good accuracy. Therefore, a good measure should consider both coverage of answer instances, denoted by *Instance Recall* ($iRecall$), and precision of returned answers ($pPrecision$). Equation 3.6 showed how $pPrecision$ can be computed. As for $iRecall$, it is similar to the *Instance Recall* measure used in evaluating list questions in the question answering tracks of TREC 2003 through 2007 [45], [46], [130], [131], [133].

$$iRecall(R) = \frac{|I_R|}{|I_A|} \tag{3.7}$$

where I_A denotes the set of *distinct* gold answer instances for the given question, which is readily constructed using the verse-to-instances map T for the question (see Figure 3.5 for an example of map T), and I_R denotes the set of *distinct* answer instances covered by the system's answers R. I_R can be constructed using the verse-to-instances map too. Finally, we compute F_1 as the performance measure of the system given the question, applied over $pPrecision$ and $iRecall$. Refer to Figure A.1 in Appendix A for an evaluation example (scenario 1) of a system's response to a multi-answer question.

3.1.6.3.2. Retrieving all occurrences of answer instances. This scenario requires systems to retrieve *all* occurrences of answer instances. The proposed evaluation measure is similar to the one introduced in the first scenario (Sec. 3.1.6.3.1) with only one modification related to $iRecall$. To construct I_A and I_R, we will consider all occurrences of instances to be distinct (i.e., treating each occurrence of an answer instance as a unique one) in both the returned and gold answers, respectively. Note that only occurrences of instances originating from different verses are to be considered distinct; therefore, if the answers comprise overlapping verses, their respective answer instances should be counted only once. This reflects the requirement of retrieving all occurrences. $pPrecision$ is computed the same way as shown earlier. Refer to the evaluation example (scenario 2) in Figure A.1 in Appendix A.

3.1.6.4. Evaluating No-Answer Questions

Although in Section 3.1.1.3, a distinction was made between zero-answer and no-direct-answer questions, we have chosen to adopt the same evaluation method for both. This is mainly attributed to the fact that *AyaTEC* does not include exhaustive sets of all indirect answers to the questions. As such, a system that does not return an answer to a zero-answer question, or only returns indirect answers to a no-direct-answer question, will be given a score of 1 for that question, and zero otherwise.

3.1.6.5. Overall Performance Evaluation

For a single figure-of-merit over the entire set of questions in *AyaTEC*, we propose to compute an overall evaluation score S for the above three question types as a weighted average of their respective scores:

$$S = w_s * S_s + w_m * S_m + w_n * S_n \tag{3.8}$$

where S_s, S_m, and S_n are the computed average scores over the single-answer, multi-answer, and no-answer question types of *AyaTEC*, respectively. The weights represent the distribution of the three question types in *AyaTEC*.

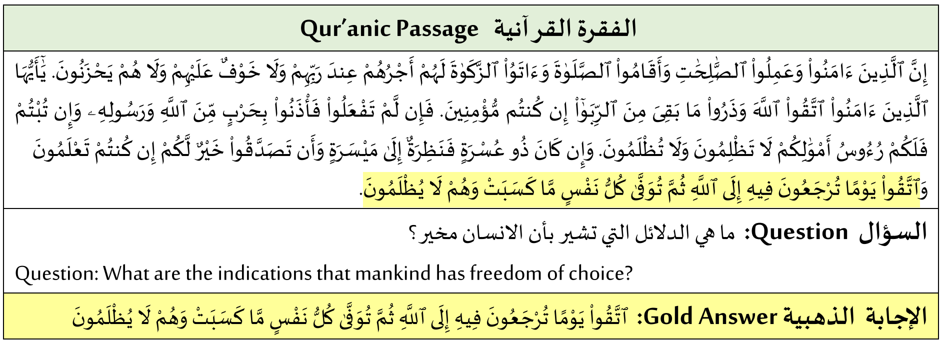
3.2. Developing QRCD: the Qur'anic Reading Comprehension Dataset

Motivated by the recent resurgence of the MRC field and its pivotal role in modern QA systems that adopt the retriever-reader architecture [38], [39], in addition to the permanent interest in Qur'an, we extend *AyaTEC* to develop *QRCD* as the first extractive Qur'anic Reading Comprehension Dataset. Extractive MRC refers to the task of span prediction, where the answer is a specific span of text extracted from passages accompanying a question [34], [38]. *QRCD* is composed of 1,337 question-passage-answer triplets for 1,093 question-passage pairs, of which 14% are multi-answer questions, which presents an additional challenge to the MRC task.

3.2.1. Extending AyaTEC *for Use in Extractive Machine Reading Comprehension*

In this section, we describe the procedure for developing *QRCD* to facilitate its use in MRC, which is currently a very active area of research. *QRCD* differs from

AyaTEC in several ways. First, it is augmented with passages curated from the Holy Qur'an to form tuples of question-passage-answer triplets adopting the same format of SQuAD v1.1 [111]. Second, the answers to the questions in *QRCD* are span-based, where the spans of text were extracted manually from their corresponding verse-based *direct* answers in AyaTEC. As such, *indirect* and *incorrect* answers were ignored. Finally, *no-answer* questions that do not have an answer in the Holy Qur'an were also ignored, keeping only the questions that have at least one answer. A *Single-answer* question is the question that has only one answer (i.e., an answer that is a single span of text, denoted as an "answer span") in the accompanying Qur'anic passage, as shown in Figure 3.8. A *multi-answer* question is the one whose answers are composed of several components (such as *list* or *why* questions) in two or more different answer spans (in distant or contiguous verses) in the accompanying Qur'anic passage, as shown in Figure 3.9.



Qur'anic Passage الفقرة القرآنية
إِنَّ ٱلَّذِينَ ءَامَنُواْ وَعَمِلُواْ ٱلصَّٰلِحَٰتِ وَأَقَامُواْ ٱلصَّلَوٰةَ وَءَاتَوُاْ ٱلزَّكَوٰةَ لَهُمْ أَجْرُهُمْ عِندَ رَبِّهِمْ وَلَا خَوْفٌ عَلَيْهِمْ وَلَا هُمْ يَحْزَنُونَ. يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُواْ ٱتَّقُواْ ٱللَّهَ وَذَرُواْ مَا بَقِيَ مِنَ ٱلرِّبَوٰٓاْ إِن كُنتُم مُّؤْمِنِينَ. فَإِن لَّمْ تَفْعَلُواْ فَأْذَنُواْ بِحَرْبٍ مِّنَ ٱللَّهِ وَرَسُولِهِۦ وَإِن تُبْتُمْ فَلَكُمْ رُءُوسُ أَمْوَٰلِكُمْ لَا تَظْلِمُونَ وَلَا تُظْلَمُونَ. وَإِن كَانَ ذُو عُسْرَةٍ فَنَظِرَةٌ إِلَىٰ مَيْسَرَةٍ وَأَن تَصَدَّقُواْ خَيْرٌ لَّكُمْ إِن كُنتُمْ تَعْلَمُونَ وَٱتَّقُواْ يَوْمًا تُرْجَعُونَ فِيهِ إِلَى ٱللَّهِ ثُمَّ تُوَفَّىٰ كُلُّ نَفْسٍ مَّا كَسَبَتْ وَهُمْ لَا يُظْلَمُونَ.
السؤال Question: ما هي الدلائل التي تشير بأن الانسان مخير؟ Question: What are the indications that mankind has freedom of choice?
الإجابة الذهبية Gold Answer: ٱتَّقُواْ يَوْمًا تُرْجَعُونَ فِيهِ إِلَى ٱللَّهِ ثُمَّ تُوَفَّىٰ كُلُّ نَفْسٍ مَّا كَسَبَتْ وَهُمْ لَا يُظْلَمُونَ

Figure 3.8. An example of a single-answer question: a single span of text.

In general, as the answer(s) to single-answer and multi-answer questions may appear in semantically and/or syntactically similar forms in different chapters and across different verses within different Qur'anic contexts, each question-passage pair in *QRCD* was considered an independent question for the MRC task. We note that each Qur'anic passage in QRCD may have more than one occurrence; and each passage occurrence

<table>
<tr><th colspan="2">Qur’anic Passage الفقرة القرآنية</th></tr>
<tr><td colspan="2">وَٱلْمُطَلَّقَٰتُ يَتَرَبَّصْنَ بِأَنفُسِهِنَّ ثَلَٰثَةَ قُرُوٓءٍ وَلَا يَحِلُّ لَهُنَّ أَن يَكْتُمْنَ مَا خَلَقَ ٱللَّهُ فِىٓ أَرْحَامِهِنَّ إِن كُنَّ يُؤْمِنَّ بِٱللَّهِ وَٱلْيَوْمِ ٱلْءَاخِرِ
وَبُعُولَتُهُنَّ أَحَقُّ بِرَدِّهِنَّ فِى ذَٰلِكَ إِنْ أَرَادُوٓا۟ إِصْلَٰحًا وَلَهُنَّ مِثْلُ ٱلَّذِى عَلَيْهِنَّ بِٱلْمَعْرُوفِ وَلِلرِّجَالِ عَلَيْهِنَّ دَرَجَةٌ وَٱللَّهُ عَزِيزٌ حَكِيمٌ.
ٱلطَّلَٰقُ مَرَّتَانِ فَإِمْسَاكٌۢ بِمَعْرُوفٍ أَوْ تَسْرِيحٌۢ بِإِحْسَٰنٍ وَلَا يَحِلُّ لَكُمْ أَن تَأْخُذُوا۟ مِمَّآ ءَاتَيْتُمُوهُنَّ شَيْـًٔا إِلَّآ أَن يَخَافَآ أَلَّا يُقِيمَا حُدُودَ
ٱللَّهِ فَإِنْ خِفْتُمْ أَلَّا يُقِيمَا حُدُودَ ٱللَّهِ فَلَا جُنَاحَ عَلَيْهِمَا فِيمَا ٱفْتَدَتْ بِهِۦ تِلْكَ حُدُودُ ٱللَّهِ فَلَا تَعْتَدُوهَا وَمَن يَتَعَدَّ حُدُودَ ٱللَّهِ
فَأُو۟لَٰٓئِكَ هُمُ ٱلظَّٰلِمُونَ. فَإِن طَلَّقَهَا فَلَا تَحِلُّ لَهُۥ مِنۢ بَعْدُ حَتَّىٰ تَنكِحَ زَوْجًا غَيْرَهُۥ فَإِن طَلَّقَهَا فَلَا جُنَاحَ عَلَيْهِمَآ أَن يَتَرَاجَعَآ إِن ظَنَّآ أَن
يُقِيمَا حُدُودَ ٱللَّهِ وَتِلْكَ حُدُودُ ٱللَّهِ يُبَيِّنُهَا لِقَوْمٍ يَعْلَمُونَ.</td></tr>
<tr><td colspan="2"><b>Question:</b> Does Islam honor women? <b>السؤال</b>: هل كرّم الإسلام المرأة؟</td></tr>
<tr><th>Predicted Answers الإجابات المسترجعة</th><th>Gold Answer الإجابات الذهبية</th></tr>
<tr><td>• ٱلطَّلَٰقُ مَرَّتَانِ فَإِمْسَاكٌۢ بِمَعْرُوفٍ أَوْ تَسْرِيحٌۢ بِإِحْسَٰنٍ
• وَلَهُنَّ مِثْلُ ٱلَّذِى عَلَيْهِنَّ بِٱلْمَعْرُوفِ وَلِلرِّجَالِ عَلَيْهِنَّ دَرَجَةٌ وَٱللَّهُ عَزِيزٌ حَكِيمٌ. ٱلطَّلَٰقُ مَرَّتَانِ فَإِمْسَاكٌۢ بِمَعْرُوفٍ أَوْ تَسْرِيحٌۢ بِإِحْسَٰنٍ
• حْسَٰنٍ
• ....</td><td>• لَهُنَّ مِثْلُ ٱلَّذِى عَلَيْهِنَّ بِٱلْمَعْرُوفِ
• ٱلطَّلَٰقُ مَرَّتَانِ فَإِمْسَاكٌۢ بِمَعْرُوفٍ أَوْ تَسْرِيحٌۢ بِإِحْسَٰنٍ
• لَا يَحِلُّ لَكُمْ أَن تَأْخُذُوا۟ مِمَّآ ءَاتَيْتُمُوهُنَّ شَيْـًٔا</td></tr>
</table>

Figure 3.9. A typical answer to a multi-answer question: two or more different spans of text, each of which is an answer component.

(if more than one exists) is paired with a different question. Likewise, each question in QRCD may have more than one occurrence; and each question occurrence (if more than one exists) is paired with a different Qur’anic passage.

Overall, *QRCD* is composed of 1,093 question-passage pairs; 939 of which are single-answer questions and the remaining 154 are multi-answer questions. With 14% of the questions in *QRCD* being multi-answer questions, this poses an additional challenge to the reading comprehension task.

3.2.1.1. Passage Curation

The Holy Qur’an is composed of 114 chapters of different lengths. We initially segmented the chapters using the Thematic Holy Qur’an [127],[7] which is a printed edition that clusters the verses of each chapter into topics. We recruited two annotators through UpWork[8] to extract the start and end verse numbers to which each topic cluster

[7] `https://surahquran.com/tafseel-quran.html`

[8] `https://www.upwork.com`

of verses starts and ends within each chapter, given the topics indicated by the printed Thematic Holy Qur'an. The text of each Qur'anic passage was then populated by appending the text of the respective verses that constitute each passage, and separating these verses by full stops. The Qur'anic text was downloaded from the Tanzil[9] project, which provides a verified digital version of the Holy Qur'an in many scripting styles in addition to the Uthmani style. We have used the normalized simple-clean text style (in Tanzil 1.0.2) to be able to use the *QRCD* dataset with transformer-based language models that were already pre-trained using normalized Arabic text. We note that Al-Azami [30] has emphasized the importance of using the Uthmani orthography when quoting or printing Qur'an verses, especially that Muslim scholars universally agree that this orthography style should be maintained.

For each Qur'anic passage, we collated all the questions of AyaTEC that have their verse-based answers fully contained within the boundaries of the passage at hand. If a verse-based answer happened to be partially contained within a Qur'anic passage, we adopted the heuristic of incrementally expanding that passage with the neighboring next verse (from the next passage) until it accommodates the full answer. Despite our effort to avoid passage overlap by adopting this expansion heuristic, some overlap in the Qur'anic passages may still exist. This segmentation procedure has resulted in 629 Qur'anic passages (associated with questions) with an average size of 80 tokens.

3.2.1.2. Answer Span Extraction

After curating the passages, we also recruited three UpWork workers (annotators), who are knowledgeable in Qur'an, to extract the specific answer spans from their respective *direct* verse-based answers given by *AyaTEC*. An interface was developed for

[9] `https://tanzil.net/download/`

that purpose, which displays a Qur'anic passage and loops over its related questions, displaying one question and its verse-based *direct* answer(s), one at a time. The annotators were *only* allowed to highlight and select the specific answer spans from the corresponding displayed *direct* verse-based answer. Each of the three annotators annotated all the questions. To resolve mismatches among extracted spans, which mostly occur due to the inclusion or exclusion of non-essential phrases, the first author resolves them. In Section 3.2.2, we further discuss the inter-annotator agreement and mismatches among the annotators.

The final number of answer spans extracted for the 1,093 questions (or question-passage pairs) was 1,337 with an average size of eight words per span. Their distribution across question types are shown in Table 3.2.

Table 3.2. Distribution of question-passage-answer triplets by question type in *QRCD*. We note that there are several untypical cases for some questions (single-answer or multi-answer), where an exact same answer may have more than one occurrence in the same Qur'anic passage.

Question Type	# Questions-Passage Pairs	# question-passage-answer triplets
Single-answer	939	949
Multi-answer	154	388
All	1093	1337

3.2.2. Inter-Annotator Agreement

As an indication of the quality of the answer span extraction phase in developing *QRCD*, we need to measure the inter-annotator agreement between our three annotators over the extracted answer spans. For that, we have adopted Fleiss Kappa [125]. We applied the measure at the *token* level. Since the annotators extracted the answers spans from the *verse-based* answers, provided in AyaTEC, rather than the whole passage [85], we computed the measure only on the tokens constituting such verses. For each token, each annotator is assigned a label of 1 or 0 based on whether the token was selected (as part of an answer span) by that annotator or not. Then, Fleiss Kappa was applied at the token level over those labels. Disagreement occurred in about 32% of the tokens, and a Kappa agreement score of 0.56 was attained. According to the Kappa interpretation scale proposed by [77], the strength of the agreement is considered *moderate*. This agreement level is similar to the one attained among the three Qur'an specialists/judges in developing AyaTEC [85].

3.2.3. Using QRCD *in Evaluation*

Performance evaluation of an extractive MRC system over a question related to a given Qur'anic passage should not be confined to one predicted answer only, especially for multi-answer questions. Therefore, we expect the *ideal* MRC system to return *all* correct answers *exclusively* (i.e., only the correct ones). Since systems are imperfect, we would like to give (partial) credit to a system that returns correct answers along with some incorrect ones; however, a system that perceives the correct answers as the *best* answers (by giving them higher scores or putting them at the top of the returned answers) should be rewarded higher than a system that perceives incorrect ones as the best. Such a system would save the user's time in checking the answers, thus better satisfying her need. This clearly calls for a *rank-based* measure, i.e., a measure that considers the ranks of the returned predicted answers. Moreover, a system that returns a partial span of a correct answer should receive a partial credit. Therefore, for our task, we need a rank-based measure that considers *partial matching* of answers. As such, we expect the system to return a *ranked list* of predicted answers R, which is evaluated against a set of one or more gold answers A to the given question. The gold answers were manually-extracted from the accompanying Qur'anic passage to that question (Section 3.2.1).

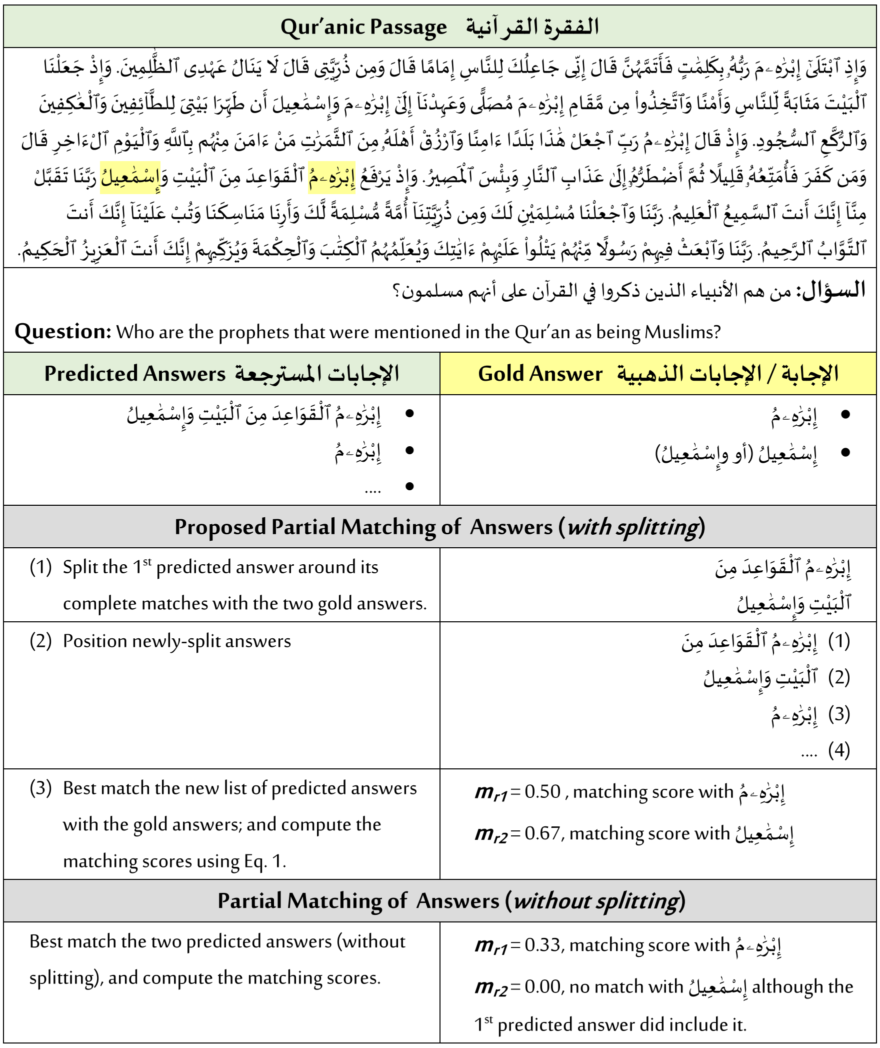
Our review of the reading comprehension literature has revealed the lack of *rank-based* evaluation measures that can integrate partial matching for evaluating extractive MRC tasks on datasets with multi-answer questions. The current evaluation measures that are being used for answer span prediction tasks mainly include the token-level F_1 (computed over bag-of-tokens) and Exact Match of answer spans (*EM*) [39], [111], [146]. While these two *set-based* measures are relatively adequate for evaluating single-answer questions, they are not adequate for multi-answer questions, because they focus the evaluation only on *one* predicted answer. Dua, Wang, Dasigi, *et al.* [49] have addressed this problem for the multi-answer questions in their DROP dataset, by extending their version of the token-level F_1 measure such that every predicted answer was best matched with one gold answer; and no gold answer was matched with more than one predicted answer for a given question. Similarly, Khashabi, Chaturvedi, Roth, *et al.* [71] also proposed an extended macro-average F_{1_m} measure for evaluating multi-answer questions. Although those two proposed F_1 measures can integrate partial matching, they are not rank-based measures; they reward the system for returning answers regardless of how they are ordered/ranked, which is not fair for systems that prefer correct answers, e.g., presenting them at the top of the returned ranked list.

Moreover, even with partial matching of answers, we need to consider cases when evaluating predicted answer spans that happen to cover more than one gold answer. With current rank-based measures, such predicted answers will be treated *unfairly*, because they will only be matched to one gold answer (at each rank) regardless of how many gold answers they may cover. Figure 3.10 exhibits an example that demonstrates such an unfair matching incidence that would cause a system to be under-evaluated. We discuss this further in the context of Section 3.2.3.1.

To address the above issues and be able to use a rank-based measure that can *fairly* integrate partial matching, we introduce a simple yet novel method to match the predicted answers against their respective gold answers (Section 3.2.3.1); and adapt the traditional Average Precision (*AP*) rank-based measure [73] to integrate *partial* matches, in addition to exact/binary matches. We denote this measure by *Partial Average Precision* (*pAP* for short), which is used as the main measure for evaluating both single-answer and multi-answer questions of the *QRCD* dataset (Sections 3.2.3.2 and 3.2.3.3 respectively).[10] The traditional *EM* and token-level F_1 evaluation measures are also adopted, but for single-answer questions only.

We note that rank-based measures were used sparingly for evaluation [34] over single-answer questions, but they were mainly applied in sentence or answer selection (rather than span extraction) tasks and without integrating partial matching [94], [134].

With the concept of partial matching with gold answers being integral to all adopted measures, we formally present it first, before defining the evaluation measures. As each measure is defined with respect to a given question, an overall evaluation score is computed by averaging over all questions, and also over questions of a specific type.



<table>
<tr><th colspan="2">Qur'anic Passage الفقرة القرآنية</th></tr>
<tr><td colspan="2">وَإِذِ ٱبْتَلَىٰٓ إِبْرَٰهِـۧمَ رَبُّهُۥ بِكَلِمَٰتٍ فَأَتَمَّهُنَّ قَالَ إِنِّى جَاعِلُكَ لِلنَّاسِ إِمَامًا قَالَ وَمِن ذُرِّيَّتِى قَالَ لَا يَنَالُ عَهْدِى ٱلظَّٰلِمِينَ. وَإِذْ جَعَلْنَا
ٱلْبَيْتَ مَثَابَةً لِّلنَّاسِ وَأَمْنًا وَٱتَّخِذُوا۟ مِن مَّقَامِ إِبْرَٰهِـۧمَ مُصَلًّى وَعَهِدْنَآ إِلَىٰٓ إِبْرَٰهِـۧمَ وَإِسْمَٰعِيلَ أَن طَهِّرَا بَيْتِىَ لِلطَّآئِفِينَ وَٱلْعَٰكِفِينَ
وَٱلرُّكَّعِ ٱلسُّجُودِ. وَإِذْ قَالَ إِبْرَٰهِـۧمُ رَبِّ ٱجْعَلْ هَٰذَا بَلَدًا ءَامِنًا وَٱرْزُقْ أَهْلَهُۥ مِنَ ٱلثَّمَرَٰتِ مَنْ ءَامَنَ مِنْهُم بِٱللَّهِ وَٱلْيَوْمِ ٱلْءَاخِرِ قَالَ
وَمَن كَفَرَ فَأُمَتِّعُهُۥ قَلِيلًا ثُمَّ أَضْطَرُّهُۥٓ إِلَىٰ عَذَابِ ٱلنَّارِ وَبِئْسَ ٱلْمَصِيرُ. وَإِذْ يَرْفَعُ إِبْرَٰهِـۧمُ ٱلْقَوَاعِدَ مِنَ ٱلْبَيْتِ وَإِسْمَٰعِيلُ رَبَّنَا تَقَبَّلْ
مِنَّآ إِنَّكَ أَنتَ ٱلسَّمِيعُ ٱلْعَلِيمُ. رَبَّنَا وَٱجْعَلْنَا مُسْلِمَيْنِ لَكَ وَمِن ذُرِّيَّتِنَآ أُمَّةً مُّسْلِمَةً لَّكَ وَأَرِنَا مَنَاسِكَنَا وَتُبْ عَلَيْنَآ إِنَّكَ أَنتَ
ٱلتَّوَّابُ ٱلرَّحِيمُ. رَبَّنَا وَٱبْعَثْ فِيهِمْ رَسُولًا مِّنْهُمْ يَتْلُوا۟ عَلَيْهِمْ ءَايَٰتِكَ وَيُعَلِّمُهُمُ ٱلْكِتَٰبَ وَٱلْحِكْمَةَ وَيُزَكِّيهِمْ إِنَّكَ أَنتَ ٱلْعَزِيزُ ٱلْحَكِيمُ.</td></tr>
<tr><td colspan="2"><b>السؤال:</b> من هم الأنبياء الذين ذكروا في القرآن على أنهم مسلمون؟
<b>Question:</b> Who are the prophets that were mentioned in the Qur'an as being Muslims?</td></tr>
<tr><th>Predicted Answers الإجابات المسترجعة</th><th>Gold Answer الإجابة / الإجابات الذهبية</th></tr>
<tr><td>• إِبْرَٰهِـۧمُ ٱلْقَوَاعِدَ مِنَ ٱلْبَيْتِ وَإِسْمَٰعِيلُ
• إِبْرَٰهِـۧمُ
• ....</td><td>• إِبْرَٰهِـۧمُ
• إِسْمَٰعِيلُ (أو وإِسْمَٰعِيلُ)</td></tr>
<tr><th colspan="2">Proposed Partial Matching of Answers (<i>with splitting</i>)</th></tr>
<tr><td>(1) Split the 1st predicted answer around its complete matches with the two gold answers.</td><td>إِبْرَٰهِـۧمُ ٱلْقَوَاعِدَ مِنَ
ٱلْبَيْتِ وَإِسْمَٰعِيلُ</td></tr>
<tr><td>(2) Position newly-split answers</td><td>(1) إِبْرَٰهِـۧمُ ٱلْقَوَاعِدَ مِنَ
(2) ٱلْبَيْتِ وَإِسْمَٰعِيلُ
(3) إِبْرَٰهِـۧمُ
(4) ....</td></tr>
<tr><td>(3) Best match the new list of predicted answers with the gold answers; and compute the matching scores using Eq. 1.</td><td>m_{r1} = 0.50 , matching score with إِبْرَٰهِـۧمُ
m_{r2} = 0.67, matching score with إِسْمَٰعِيلُ</td></tr>
<tr><th colspan="2">Partial Matching of Answers (<i>without splitting</i>)</th></tr>
<tr><td>Best match the two predicted answers (without splitting), and compute the matching scores.</td><td>m_{r1} = 0.33, matching score with إِبْرَٰهِـۧمُ
m_{r2} = 0.00, no match with إِسْمَٰعِيلُ although the 1st predicted answer did include it.</td></tr>
</table>

Figure 3.10. An example that compares the proposed partial matching of answers (with splitting), to the traditional partial matching (without splitting), and their implications on the computed matching scores that would unfairly cause a system to be under-evaluated.

3.2.3.1. Partial Matching of Answers

Reading comprehension systems might predict answers that are not *exact* matches to any of the gold answers for a given question, despite matching it partially, or even covering it completely within a larger span. To give partial and fair credit to such systems, we start the matching process by computing the *span overlap* between every system's predicted answer and all the gold answers that it overlaps with partially or fully.

[10] Other rank-based measures, such as *nDCG* can also be adapted.

In case a predicted answer matches (i.e., overlaps with) more than one gold answer, it is then *split* around its respective matches with the gold answers. In that case, the newly-split answers will *replace* the original answer in the ranked list, with the same order they appear in the original answer. Naturally, no splitting is applied if the predicted answer does not match any gold answer, or if it includes a match (partial or full) with only one gold answer. Finally, every answer in the newly-formed (expanded) ranked list of predicted answers is best matched with one gold answer. Henceforward, we refer to the proposed matching method as partial matching *with splitting*, as opposed to the traditional partial matching *without splitting*. An example of the proposed answer matching is presented in Figure 3.10.

We note that partial matching *with splitting* induces a ripple effect on the rank order of subsequent predicted answers (as shown in Figure 3.10), which will, in turn, have a direct effect on the computation of our proposed rank-based measure. It is worth emphasizing that splitting is performed *only* to address cases when one predicted answer matches *more than one* gold answer. If the traditional matching (*without splitting*) is used, that predicted answer would match only one gold answer, which would be unfair (as clearly shown in Figure 3.10). However, splitting allows giving credit for matching all of those gold answers. The only side effect is the increase/expansion of the ranked list. We note that this is quite natural, as it follows the sequential order of reading the words of the predicted answer, matching the incremental perceived gain in user satisfaction when reading the correct answers sequentially within the words of the predicted answer.

We have adopted the definition by Malhas and Elsayed [85] for the answer matching score m of a system's predicted answer r, which was denoted by m_r. It was defined as the maximum matching score of answer r over all the gold answers A for a given question, such that each best matched gold answer can only be matched once.

$$m_r = \max_{a \in A} F_1(r, a) \tag{3.9}$$

where F_1 is computed here over token *positions*, rather than any arbitrary matching bag-of-tokens, to reward a predicted answer only if it was extracted from the proper verse/context. Figure 3.10 compares the answer matching scores computed based on the proposed and traditional matching methods (i.e., with or without splitting), to demonstrate how our proposed matching avoids the unfair deterioration of the scores computed using the traditional matching method.

3.2.3.2. Evaluating Single-Answer Questions

The first two evaluation measures that we have adopted for single-answer questions were F_1 and *EM*, which were both applied by Rajpurkar, Zhang, Lopyrev, *et al.* [111] on the top predicted answer against its ground truth answer.

We use the term $F_1@1$ to refer to F_1 when applied on the predicted answer at the *first* rank only.

$$F_1@1(R) = m_{r_1} \tag{3.10}$$

where R is the system's returned ranked list of predicted answers, and r_1 is the predicted answer at the first rank in R.

We also use *EM*, which is a binary measure that checks whether the first predicted answer *exactly matches* the gold answer to a given question. We formally define *EM* in terms of the answer matching score at the first rank m_{r_1}.

$$EM(R) = \begin{cases} 1 & \text{if } m_{r_1} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

The third adopted measure *pAP* is described in the next section as it is also used for evaluating multi-answer questions.

3.2.3.3. Evaluating Multi-Answer Questions

The F_1 (or $F_1@1$) and *EM* measures are not suitable for evaluating multi-answer questions because they only focus on the top predicted answer, ignoring the others. Moreover, with the task being perceived as a ranking problem, it is important to adopt a *rank-based* measure that can also assess partial matches. As such, we introduce *Partial Average Precision* (*pAP*) as a variant of the traditional Average Precision (*AP*) rank-based measure, to integrate the concept of partial matching, and use it to evaluate multi-answer as well as single-answer questions.[11] *pAP* is defined as follows.

$$pAP(R) = \frac{1}{|A|} \sum_{K=1}^{|R|} \mathbb{1}\{m_{r_K} > 0\} \cdot pPrec@K(R) \tag{3.12}$$

where $|R|$ and $|A|$ are the number of answers in the system's returned ranked list R and the gold answers A, respectively, r_K is the predicted answer at the rank K in R, and $\mathbb{1}\{m_{r_K} > 0)\}$ is the indicator function that has a value of 1 only if the predicted answer at rank K matches (partially or fully) a gold answer, and zero otherwise. *Partial Precision at rank K, denoted as* pPrec@K, *is a variant of the traditional* Prec@K *measure that*

[11] Similar to the traditional Average Precision (*AP*) [73], *pAP* averages the computed (here partial) precision at the ranks of each predicted answer that (partially or fully) matches a gold answer (assuming that non-retrieved gold answers appear at very low rank for which precision is zero).

also integrates the concept of partial matching, defined by Malhas and Elsayed [85] as follows:

$$pPrec@K(R) = \frac{1}{K}\sum_{i=1}^{K} m_{r_i} \tag{3.13}$$

where R is the system's returned ranked list of predicted answers, r_i is the predicted answer at rank i in R, and m_{r_i} is the partial matching score of r_i as defined by Equation 3.9.

To elaborate more on how the pAP measure is computed and showcase its fairness, Figure B.1 in Appendix B presents a detailed example for the performance evaluation of the output of two different systems on one question using pAP. Although both systems predict the same set of answers, pAP better rewards the first system over the second, because it predicts the correct answers at ranks 1 and 2, while the second predicts them at lower ranks down the list.

We note that despite the change in rank order that may be induced due to partial matching *with splitting*, the gains in the matching score values are expected to outweigh any deterioration of $pPrec@K(R)$ due to the expanded rank order, as discussed in Section 3.2.3.1.

We note that all of the above measures are applied to the predicted answers for one given question.

3.2.3.4. Implications for Using QRCD in Evaluation

Having proposed the evaluation measures to use for single-answer questions and multi-answer questions, it is important to note that the question types and the scope of the evaluation have implications on the proposed evaluation measures in the previous two sections. To be more specific, we define a passage-scope and a Qur'an-scope for the evaluation. The *passage-scope* is confined to the passage accompanying the question to which its answer(s) were extracted from, while the *Qur'an-scope* comprise the whole Qur'an given that the answer(s) to a given question (of type single-answer or multi-answer) may appear in semantically and/or syntactically similar forms in different chapters and across different verses within different Qur'anic contexts.

Based on the forgoing, the passage-scope is adopted for evaluating the reader component, and the Qur'an-scope is adopted for evaluating the retriever component and the pipelined end-to-end QA system. However, this implies that a multi-answer question with two or more answer components (i.e., answer spans) in the Qur'an, will be evaluated as a multi-answer question in the Qur'an-scope evaluation of the end-to-end QA system; and it may be evaluated as a single-answer question in the passage-scope evaluation of the reader component, if the question happens to be coupled with a Qur'anic passage comprising only *one* of the question's answer components. As such, the adopted scope will also influence whether the question is classified as single-answer or multi-answer.

CHAPTER 4: THE RETRIEVER - PASSAGE RETRIEVAL WITH DOCUMENT EXPANSION

In this chapter, we describe our sparse retrieval approach towards the development of the retriever component of our QA system (Figure 1.5). Given a question in MSA, the retriever should retrieve the top k answer-bearing passages from the Holy Qur'an. These passages are then passed to the reader for answer extraction.

In general, sparse retrieval refers to traditional IR methods that use sparse bag-of-words text representation approaches to measure term overlap. In addition to the classical challenge of vocabulary mismatch that any retrieval system should overcome, our retriever component needs to address the challenge of how to bridge the gap between a question posed in MSA and the Qur'anic answer-bearing passages to be retrieved. Although MSA and CA share the same morphology and syntax characteristics, they mainly differ in lexis, where contemporary western words found their way into MSA through translation or transliteration and obsolete words were dropped [98]. Nevertheless, CA remains richer in lexis [121], especially when the Classical Arabic text is Qur'anic. With the digital presence of ample Qur'an related resources in MSA on the Web, we resorted to document expansion rather than query expansion to mitigate this gap. As such, two important and widely used MSA resources were selected; the first resource is Al-Tafseer Al-Muyassar [1], which is a simple interpretation of the Holy Qur'an in MSA, while the second is Kalimat Al-Qur'an [84], which is a dictionary of Qur'anic words with their corresponding meaning in MSA.

Figure 4.1 exhibits the adopted pipelined methodology in developing the retriever component; 1) segmenting the 114 chapters of the Qur'an into topical passages to constitute our Qur'anic passage collection; 2) cleaning and preprocessing the scrapped

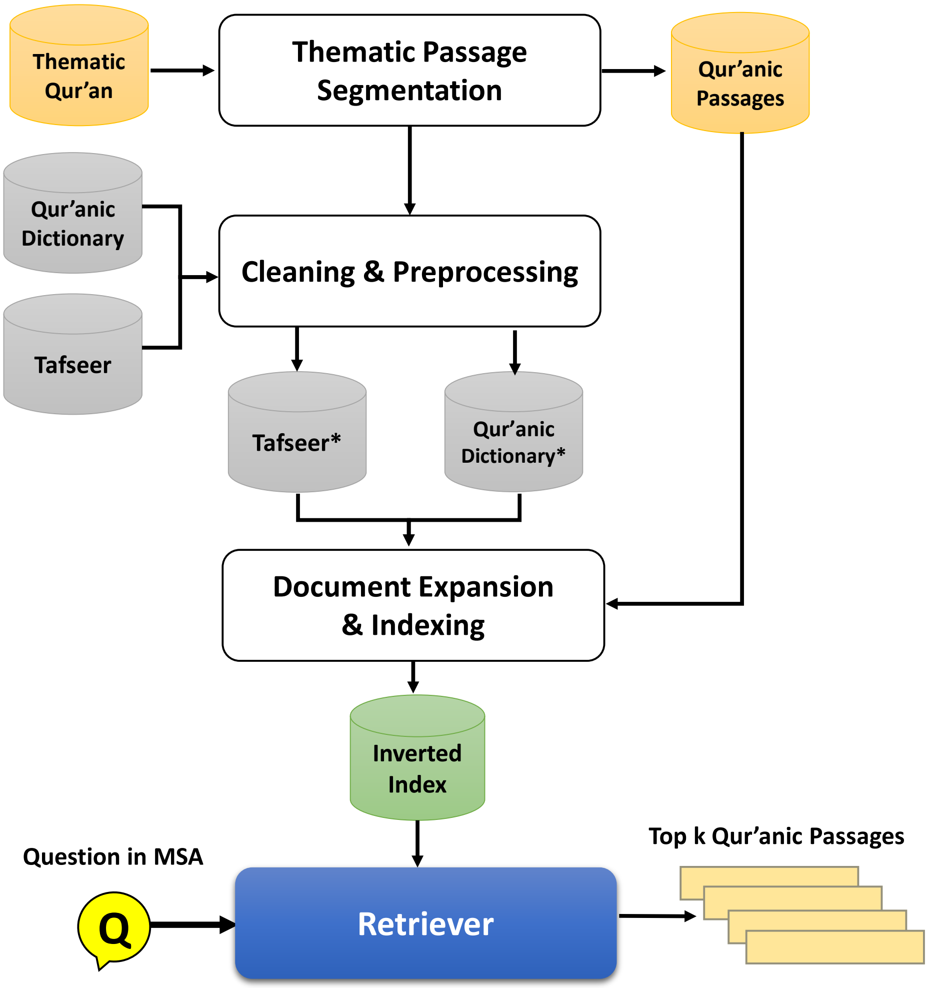
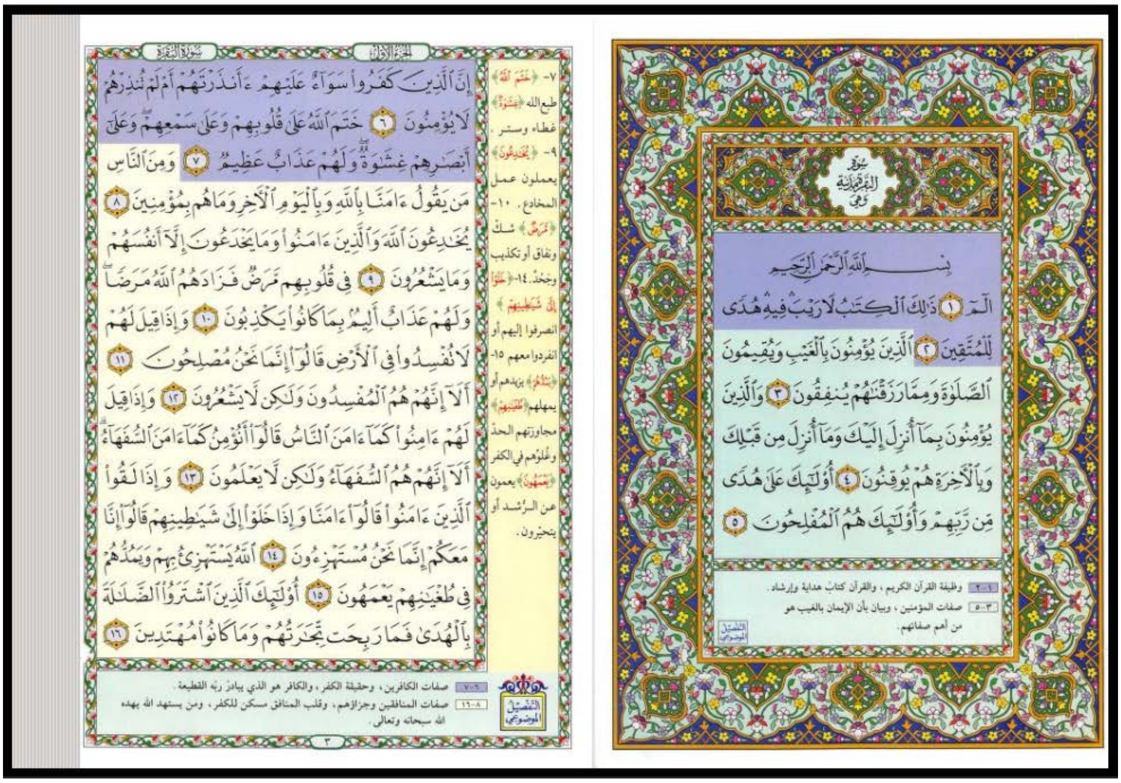


Figure 4.1. The pipelined methodology adopted for developing the Retriever component.

versions of Al-Tafseer Al-Muyassar [1] (or Al-Tafseer for short) and Kalimat Al-Qur’an dictionary [84] (or dictionary for short); 3) expanding the passages in the collection with their corresponding interpretation from Al-Tafseer, and the meanings (in MSA) of the corresponding Qur’anic words, if they exist in the dictionary; 4) indexing the expanded Qur’anic passage collections using the Pyserini tool [81]; and finally 5) searching the indexes to retrieve relevant Qur’anic passages with respect to a given question. In Sections 4.1 through 4.4, we describe each of the above steps in more detail. Then, we describe the approach adopted for developing the relevance judgments to the questions in *QRCD* to evaluate the performance of the retriever over the holdout set from *QRCD*.

We conclude this chapter with a discussion of the results to answer the first research question in this dissertation.

4.1. Thematic Passage Segmentation



سُورَةُ البَقَرَةِ

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ
الٓمٓ (١) ذَٰلِكَ ٱلْكِتَٰبُ لَا رَيْبَ فِيهِ هُدًى
لِّلْمُتَّقِينَ (٢) ٱلَّذِينَ يُؤْمِنُونَ بِٱلْغَيْبِ وَيُقِيمُونَ
ٱلصَّلَوٰةَ وَمِمَّا رَزَقْنَٰهُمْ يُنفِقُونَ (٣) وَٱلَّذِينَ
يُؤْمِنُونَ بِمَآ أُنزِلَ إِلَيْكَ وَمَآ أُنزِلَ مِن قَبْلِكَ
وَبِٱلْءَاخِرَةِ هُمْ يُوقِنُونَ (٤) أُوْلَٰٓئِكَ عَلَىٰ هُدًى
مِّن رَّبِّهِمْ وَأُوْلَٰٓئِكَ هُمُ ٱلْمُفْلِحُونَ (٥)

١-٢ وظيفة القرآن الكريم، والقرآن كتابُ هداية وإرشاد.
٣-٥ صفات المؤمنين، وبيان بأن الإيمان بالغيب هو
من أهم صفاتهم.

التفصيل الموضوعي

الجزء الأول — سورة البقرة

إِنَّ ٱلَّذِينَ كَفَرُواْ سَوَآءٌ عَلَيْهِمْ ءَأَنذَرْتَهُمْ أَمْ لَمْ تُنذِرْهُمْ
لَا يُؤْمِنُونَ (٦) خَتَمَ ٱللَّهُ عَلَىٰ قُلُوبِهِمْ وَعَلَىٰ سَمْعِهِمْ وَعَلَىٰٓ
أَبْصَٰرِهِمْ غِشَٰوَةٌ وَلَهُمْ عَذَابٌ عَظِيمٌ (٧) وَمِنَ ٱلنَّاسِ
مَن يَقُولُ ءَامَنَّا بِٱللَّهِ وَبِٱلْيَوْمِ ٱلْءَاخِرِ وَمَا هُم بِمُؤْمِنِينَ (٨)
يُخَٰدِعُونَ ٱللَّهَ وَٱلَّذِينَ ءَامَنُواْ وَمَا يَخْدَعُونَ إِلَّآ أَنفُسَهُمْ
وَمَا يَشْعُرُونَ (٩) فِي قُلُوبِهِم مَّرَضٌ فَزَادَهُمُ ٱللَّهُ مَرَضًا
وَلَهُمْ عَذَابٌ أَلِيمٌۢ بِمَا كَانُواْ يَكْذِبُونَ (١٠) وَإِذَا قِيلَ لَهُمْ
لَا تُفْسِدُواْ فِي ٱلْأَرْضِ قَالُوٓاْ إِنَّمَا نَحْنُ مُصْلِحُونَ (١١)
أَلَآ إِنَّهُمْ هُمُ ٱلْمُفْسِدُونَ وَلَٰكِن لَّا يَشْعُرُونَ (١٢) وَإِذَا قِيلَ
لَهُمْ ءَامِنُواْ كَمَآ ءَامَنَ ٱلنَّاسُ قَالُوٓاْ أَنُؤْمِنُ كَمَآ ءَامَنَ ٱلسُّفَهَآءُ
أَلَآ إِنَّهُمْ هُمُ ٱلسُّفَهَآءُ وَلَٰكِن لَّا يَعْلَمُونَ (١٣) وَإِذَا لَقُواْ
ٱلَّذِينَ ءَامَنُواْ قَالُوٓاْ ءَامَنَّا وَإِذَا خَلَوْاْ إِلَىٰ شَيَٰطِينِهِمْ قَالُوٓاْ إِنَّا
مَعَكُمْ إِنَّمَا نَحْنُ مُسْتَهْزِءُونَ (١٤) ٱللَّهُ يَسْتَهْزِئُ بِهِمْ وَيَمُدُّهُمْ
فِي طُغْيَٰنِهِمْ يَعْمَهُونَ (١٥) أُوْلَٰٓئِكَ ٱلَّذِينَ ٱشْتَرَوُاْ ٱلضَّلَٰلَةَ
بِٱلْهُدَىٰ فَمَا رَبِحَت تِّجَٰرَتُهُمْ وَمَا كَانُواْ مُهْتَدِينَ (١٦)

٧- ﴿خَتَمَ ٱللَّهُ﴾ طبع الله ﴿غِشَٰوَةٌ﴾ غطاء وستر. ٩- ﴿يُخَٰدِعُونَ﴾ يعملون عمل المخادع. ١٠- ﴿مَّرَضٌ﴾ شك ونفاق أو تكذيب وجحد. ١٤- ﴿خَلَوْاْ إِلَىٰ شَيَٰطِينِهِمْ﴾ انصرفوا إليهم أو انفردوا معهم ١٥- ﴿يَمُدُّهُمْ﴾ يزيدهم أو يمهلهم ﴿طُغْيَٰنِهِمْ﴾ مجاوزتهم الحدّ وغلوّهم في الكفر ﴿يَعْمَهُونَ﴾ يعمون عن الرُّشد أو يتحيرون.

٦-٧ صفات الكافرين، وحقيقة الكفر، والكافر هو الذي يبادرُ ربّه القطيعة.
٨-١٦ صفات المنافقين وجزاؤهم، وقلب المنافق مسكن للكفر، ومن يستهد الله يهده
الله سبحانه وتعالى.

التفصيل الموضوعي

٣

Figure 4.2. Two pages from the Thematic Holy Qur'an categorized into different themes by color. The description and range (designated by the start and end verse numbers) of each theme/topic are provided at the bottom of the pages to which their respective ranges start.

Passage segmentation of the 114 chapters of the Holy Qur'an is a step that was already conducted during the passage curation phase of developing *QRCD* as described in Section 3.2.1.1. The Thematic Holy Qur'an [127][1] was used in the segmentation to generate topical Qur'anic passages that constitute our Qur'anic document (or more precisely passage) collection of the Holy Qur'an. Figure 4.2 exhibits two pages from the Thematic Qur'an segmented into four themes/topics using different colors, to visually

[1] https://surahquran.com/tafseel-quran.html

separate the topical clusters in those pages according to their respective descriptions. We note that some of the Qur'anic passages in *QRCD* may not fully correspond to the boundaries of this Qur'anic passage collection.[2]

4.2. Cleaning and Preprocessing the Data used in Document Expansion

After scrapping the digital content of Al-Tafsser Al-Muyassar [1][3] and Kalimat Al-Qur'an dictionary [84],[4] two UpWork workers were hired to manually clean the data of the Qur'an dictionary. Henceforward, we will refer to these two resources jointly as "MSA resources for expansion". As a preprocessing step, we normalized the MSA text in each resource such that it complies with the simple-clean text style in Tanzil 1.0.2[5] of the Qur'anic text that we have downloaded and used while curating the Qur'anic passages in *QRCD* (Section 3.2.1.1).

4.3. Expanding Passages with MSA Rsources

To mitigate the gap between the MSA questions and their Qur'anic passages answer-bearing, we start by adopting the listed below expansion alternatives to create the expanded versions of our Qur'anic passage collection; each Qur'anic passage (i.e., set of verses) was expanded with the corresponding MSA text relevant to the same verses that constitute that passage.

1. Expansion with Kalimat Al-Qur'an dictionary. Only the meanings of the Qur'anic words in MSA were used in the expansion, if they exist, as not all

[2] Due to applying some expansion heuristics to accommodate the full direct answers (i.e, qur'anic verses) from *AyaTEC* prior to span extraction, some of the Qur'anic passages in *QRCD* were merged or incrementally expanded.

[3] Al-Tafsser was scrapped from `https://quranenc.com/ar/browse/arabic_moyassar/`

[4] Dictionary was scrapped from `https://www.e-quran.com/indx-word.html`

[5] `https://tanzil.net/download/`

Qur'anic words have meanings in the dictionary.

2. Expansion with Al-Tafeer Al-Muyassar only.

3. Expansion with both, Al-Tafseer and the dictionary.

4.4. Indexing and Searching

We indexed the four topical Qur'anic passage collections (including the non-expanded version of the collection, which is the baseline) using the Pyserini tool [81].

Given a question in MSA, the retriever searches any of the indexes using the Okapi BM25 [113] scoring function to retrieve the top k answer-bearing Qur'anic passages for that question, which are then passed to the reader as pure Qur'anic passages without the MSA text.

4.5. Relevance Judgments

The relevance judgements of the questions in *QRCD* over the Qur'anic passage collection were created using the respective gold answers of those questions. Each Qur'anic passage in the collection was considered relevant to the question, if it happened to comprise any of the gold answer(s) completely or partially.

4.6. Experimental Evaluation of the Retriever

In this section, we describe the setup of our experiments, then present the evaluation results and discuss them in the context of addressing the first research question:

RQ1: Would expanding the Qur'anic passages with their corresponding Qur'an related MSA resources help the retriever in bridging the gap between the questions in

MSA and their answer-bearing Qur'anic passages?

Table 4.1. Distribution of questions and their question-passage-answer triplets in QRCD. The distribution and the counts are based on the Qur'an-scope.

Dataset	# Questions	# Question-passage-answer triplets		
		All Questions	Single-answer Questions	Multi-answer Questions
All	169	1337	44	1293
Training	135	989	31	958
Test / Holdout	34	348	13	335

For evaluating the retriever, we randomly split the unique questions in *QRCD* (i.e., without their accompanying passages) into a holdout (20%) dataset and a training (80%) dataset. The holdout dataset is composed of 34 questions (as shown in Table 4.1). Adopting the Qur'an-scope for evaluation (Section 3.2.3.4), we opted to use Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) in evaluating the retriever, and report recall at different ranks (Table 4.2).

4.6.1. Results and Discussion

Table 4.2 presents the evaluation results of the retriever over the different indexes. To address RQ1 that is concerned with observing the effect of expanding the Qur'anic passage collection with MSA resources in bridging the gap between MSA questions and their Qur'anic answer-bearing passages, we evaluate the performance of the retriever over the non-expanded Qur'anic collection (as the baseline), and over the indexes of the three expanded Qur'anic passage collections with variant combinations of the two MSA resources (Section 4.3).

Table 4.2. Performance evaluation of the retriever over the indexes of the baseline and expanded Qur'anic passage collections. Best evaluation scores are boldfaced per measure.

Qur'anic Passage Collection	MAP	MRR	recall@5	recall@10	recall@20	recall@30
No expansion - Qur'an only (baseline)	20.79	30.71	19.13	29.52	35.03	43.10
Expanded with Al-Qur'an Dictionary	20.18	32.16	21.46	25.18	36.72	43.47
Expanded with Al-Tafseer Al-Muyassar	30.58	44.94	35.82	**49.35**	53.28	**59.24**
Expanded with Al-Tafseer and Dictionary	**30.84**	**45.77**	**36.31**	48.44	**53.30**	59.09

The results reveal two important findings. First, passage expansion using Al-Tafseer Al-Muyassar is the most effective. This is evident from the comparable evaluation scores attained by the retriever over the two indexes that have Al-Tafseer used in expansion, regardless of whether the Qur'an dictionary was used or not (without overlooking the marginal increase in scores due to the dictionary). Second, retrieval over the index expanded with both MSA resources attained an increase of 10.1 and 15.1 points on its MAP and MRR scores, respectively, in comparison to the baseline, and an increase of 17+ points on its recall scores at ranks 10, 20 and 30.

As such, to answer RQ1, the results demonstrate the effective contribution of passage expansion with MSA resources in mitigating the gap between the MSA questions and their Qur'anic answer-bearing passages. Nevertheless, the performance of our best performing retriever is relatively modest. Inspecting some of the failure examples revealed the need for semantic similarity approaches to down-weigh or prohibit the retrieval of passages with high term overlap that do not contain an answer, or contain the right answer but in a disparate context (i.e., hard negatives). We believe the retriever component can be improved by re-ranking the retrieved passages, and/or using dense (embedding-based) retrieval approaches that leverage semantic similarity; thus, over-coming the limitations of sparse bag-of-words (keyword-based) retrieval approaches.

CHAPTER 5: THE READER - A MACHINE READING COMPREHENSION SYSTEM ON THE HOLY QUR’AN

In this chapter, we describe our approach in developing the reader component of the QA system (Figure 1.5), Given a question-passage pair, the reader should extract the best answer(s) to the given MSA question from the accompanying Qur’anic passage.

After being dormant for decades, the MRC field witnessed a resurgence that was mainly attributed to the development of large reading comprehension datasets [67], [75], [111], which enabled the training of deep learning neural MRC systems. Moreover, the phenomenal success of transformer-based pre-trained language models, e.g., BERT [48], RoBERTa [83] and XLNet [141], have further escalated the rate at which the field of neural MRC was progressing. Interestingly, the perception towards the task has evolved from being a question answering (QA) task over a closed piece of text into an integral component of modern AI systems, such as machine reading at scale systems that adopt the Retriever-Reader architecture [39], [40], [140], [148], which we have also adopted for developing our closed-domain QA system on the Holy Qur’an.

This chapter is composed of two main parts. The sections in the first part cover the procedure of further pre-training an AraBERT-based [23] model using a Classical Arabic dataset, prior to applying a pipelined fine-tuning procedure using two MSA MRC datasets, in addition to our *QRCD* dataset to constitute our CL-AraBERT reader. The sections in the second part of this chapter are dedicated to the experimental evaluation of the CL-AraBERT reader, where we describe the experimental setup, then present the evaluation results and discuss them and their implications in the context of addressing three of the research questions in this dissertation.

Unsupervised transfer learning through pre-trained language models (LM) for text representation has been proven to be very effective in advancing various NLP tasks, especially for low-resourced languages [48]. This is mainly attributed to the unsupervised (or self-supervised) nature of LM pre-training, the ubiquitous presence of unlabelled text to train on, and the advent of transformer-based models such as GPT [109] and BERT [48] among others.

For our reading comprehension task on the Holy Qur'an, we note that the document collection of *QRCD* is in Classical Arabic (CA), whereas the questions are expressed in Modern Standard Arabic (MSA). This allows us to cast our task as a supervised cross-lingual transfer task, where the question is in one language (MSA) and the context/passage (from which the answer(s) are extracted) is in another language (CA).

Although there are some similarities between CA and MSA, CA is relatively different; therefore we expect that a language model that is pre-trained in CA will be a better fit for our purpose than a language model that is pre-trained in MSA (i.e., using MSA resources only), such as AraBERT [23]. To achieve that, we have adapted AraBERT by further pre-training it using CA resources to introduce **CL-AraBERT**. Our decision not to pre-train a BERT model from scratch using CA resources only, was driven by two factors: (i) to achieve a better cross-lingual transfer between MSA and CA, as the questions are in MSA; and (ii) to exploit the existing similarity between MSA and CA with respect to morphology and syntax characteristics. To adapt CL-AraBERT for our reading comprehension task, we then fine-tune it as a reader using two MRC datasets in MSA by Mozannar, Maamary, El Hajal, *et al.* [97], prior to further fine-tuning the reader model using the *QRCD* dataset. As such, we have overcome the lack of MRC datasets in CA and the modest size of *QRCD*, and more importantly, attempted to bridge the gap between the questions being in MSA and the answers being in Qur'anic CA.

For developing CL-AraBERT, we have followed the same pre-training and fine-tuning procedures adopted in developing BERT [48] and AraBERT models. In Section 5.1.1, we describe the pre-training dataset and the cleaning and pre-processing procedures adopted. This is followed by a detailed description of the pre-training and fine-tuning procedures of CL-AraBERT in Sections 5.1.2 and 5.1.3, respectively.

5.1.1. Classical Arabic Data for Pre-training

Devlin, Chang, Lee, *et al.* [48] have primarily released pre-trained monolingual BERT models for the English and Chinese languages, in addition to a multilingual model (mBERT) that was pre-trained using more than 100 languages, among which was the Arabic language. With the limited data and vocabulary representation for Arabic in multilingual BERT, Antoun, Baly, and Hajj [23] introduced AraBERT by pre-training a monolingual BERT model for the Arabic language using two publicly available large Arabic news corpora: (i) the Arabic Corpus of 1.5 billion words by El-Khair [70], and (ii) the OSIAN corpus by Zeroual, Goldhahn, Eckart, *et al.* [147]. As such, all their pre-training data resources were in MSA. The size of their final pre-training dataset was $\sim$24GB with about 3B words. Two versions of AraBERT were released, AraBERTv0.1 and AraBERTv1. The main difference between the two versions is that the words of the dataset used to pre-train AraBERTv1 were segmented using the Farasa tool [3] into stems, prefixes and suffixes. After learning the vocabulary using a BERT-compatible tokenizer, the final size of the vocabulary amounted to 64k tokens for both, AraBERTv0.1 and AraBERTv1, of which 4k tokens were unused to cater for learning additional tokens if further pre-training is to be conducted [23]. We have chosen to use AraBERTv0.1.

As AraBERT was pre-trained using MSA resources only, we used the OpenITI corpus [114] as the main resource for Classical Arabic to further pre-train AraBERT; we called the adapted model CL-AraBERT. We have used the OpenITI version 2019.1.1,[1] which is a machine-readable historical corpus of Arabic texts written between the years 1-1340 Hijri. We selected Arabic texts from two of OpenITI's main sources; namely, Al-Maktaba Al-Shamela[2] and Al-Jami' Al-Kabir,[3] both of which are large digital libraries of pre-modern and modern Arabic texts. The texts span a wide range of genres including Tafseer (Qur'an exegesis), Hadith, Fiqh (Islamic jurisprudence), Aqeedah (creed), literature, poetry, among others.

Extensive cleaning and preprocessing was conducted on the selected OpenITI documents because we used a raw version of the OpenITI v2019.1.1 text, which was tagged using OpenITI mARkdown.[4] It is a simple system for tagging structural, morphological, and semantic elements embedded in the OpenITI text. We also applied the same preprocessing adopted by AraBERT. The final size of the pre-training dataset amounted to about 1.05B words.

5.1.2. Pre-training CL-AraBERT

We followed the same pre-training setup and procedure adopted for building BERT$_{\text{BASE}}$. The model architecture is composed of 12 transformer layers/blocks, a hidden size of 768, and 12 self-attention heads with a total of 110M parameters to

[1] https://zenodo.org/record/3082464#.YQR_Y44zaMo

[2] https://shamela.ws/

[3] According to this link https://alraqmiyyat.github.io/OpenITI/, texts coming from Al-Jami' Al-Kabir have been published on an external HDD and are not available online. The meta data at the beginning of each document in the OpenITI corpus explicitly specifies the source from which it was obtained.

[4] https://maximromanov.github.io/mARkdown/

further pre-train.

With the OpenITI pre-training dataset ready, the next step was to use it to learn the vocabulary of the CL-AraBERT model using a tokenizer that is compatible with the WordPiece tokenizer[5] used in BERT to learn the vocabulary and generate the WordPiece embeddings [137]. We applied the Hugging Face implementation of the BERT WordPiece tokenizer. The new vocabulary was then merged (excluding duplicates) with the original vocabulary that was initially published with AraBERTv0.1,[6] such that the new vocab tokens replaced [UNUSED] placeholder tokens. The total number of vocab tokens remained at 64k.

Naturally, we adopted the same input representations and definitions used by BERT/AraBERT. In [48], a "sentence" was defined as any span of consecutive text (rather than a usual linguistic sentence), and a "sequence" was defined as the input token sequence to BERT. We constructed each input sequence by packing the WordPiece tokens of pairs of sentences (A and B) selected from the pre-training dataset as one single sequence, which we separate by the special [SEP] token. In addition, a [CLS] token and another [SEP] token were concatenated to the beginning and end of the input sequence, respectively. Then the learned embeddings for each sentence were added to the respective tokens in the input sequence. Lastly, learned position embeddings that represent the position of the token in the input sequence was added to each token. As such, the input representation of each token was constructed by adding up three embeddings, the WordPiece token embedding, the sentence embedding that the token belongs to, and the position embedding.

Starting from the trained checkpoints of AraBERTv0.1, we further pre-trained

[5] https://github.com/huggingface/tokenizers/tree/master/bindings/python/py_src/tokenizers/implementations

[6] https://github.com/aub-mind/arabert/tree/master/arabert

the model using two unsupervised tasks: the *Masked Language Model* task (MLM), and the *Next Sentence Prediction* (NSP) task. Both tasks were applied following the same procedure in BERT/AraBERT.

The MLM task was applied by randomly masking 15% of the WordPiece tokens in the input sequence to AraBERT. In this way, bidirectional learning was enforced because the objective is to predict the original vocabulary id of the masked token conditioned on its left and right contexts. It is important to note that masking of tokens happens only during pre-training and not during fine-tuning, which may create a mismatch because the [MASK] token is only seen during pre-training and never during fine-tuning. To alleviate the effect of this mismatch, a heuristic was adopted to have the training data generator replace the masked tokens with: (i) any random token 10% of the time, (ii) the original token 10% of the time, and (iii) the [MASK] token 80% of the time [23], [48].

As for the NSP task, the training examples were trivially constructed by randomly selecting and pairing two consecutive sentences as positive examples 50% of the time, and non-consecutive sentences as negative examples for the remaining 50%. The importance of the *next sentence prediction* task lies in training the model to identify relationships between sentences, which is especially important for downstream tasks such as question answering and natural language inference [23], [48].

We pre-trained CL-AraBERT on a cloud TPUv3-8 for 440k steps, which is approximately equivalent to 27 epochs over the pre-training dataset of $\sim$1.05B words. For the first 315k steps, we trained on input sequences of 128 tokens with a batch size of 512 examples. As for the remaining 125k steps, we trained on input sequences of 512 tokens with a batch size of 128 examples. The random seed and duplication factor were kept at 34 and 10, respectively (as set by Antoun, Baly, and Hajj). We used Adam with a learning rate of 2e-5, as opposed to the smaller learning rate of 1e-4 used to pre-train AraBERT from scratch.[7] Transforming the sharded pre-training dataset into TFRecords consumed 44 hours on a virtual machine with 8 vCPUs and 52 GB memory, while pre-training CL-AraBERT consumed ∼29 hours on the cloud TPU.

5.1.3. Fine-tuning CL-AraBERT

As the questions in *QRCD* include multi-answer questions that typically have two or more answer components, each of which constitutes a different answer span from the same passage, we formulate the span prediction task as a ranking problem. The reader should return a list of the best-predicted answers or answer components ranked by their probability scores.

Since the size of *QRCD* is relatively modest (Table 3.2), we leverage cross-lingual transfer learning by using the Arabic SQuAD and ARCD question answering datasets by Mozannar, Maamary, El Hajal, *et al.* [97] in fine-tuning CL-AraBERT, prior to fine-tuning the model using *QRCD*. The Arabic SQuAD is a Google translated segment of the English SQuAD v1.1 dataset to Arabic (in MSA); it comprises 48.3k QA pairs that were translated with their corresponding articles. The ARCD dataset is composed of 1,395 question-passage-answer tuples in MSA as well; we only used the training split of the dataset for training (695 tuples).

The input representation for fine-tuning is very similar to pre-training, where the tokens of each question and passage are packed as one single sequence separated by the [SEP] token. A [CLS] token and another [SEP] token are also concatenated to the beginning and end of the sequence, respectively. Similar to pre-training, the input

[7] `https://github.com/google-research/bert#pre-training-tips-and-caveats`

representation of each token was constructed by adding up its WordPiece embedding, the question or passage embedding that the token belongs to, and finally the token's position embedding.

Fine-tuning was effected by introducing two vectors, a start vector S and an end vector E. To find the best prediction for an answer span, the probability of a word i being the start of the answer span was computed as the dot product between the start vector S and the output token embedding for the word i (as captured from the last transformer hidden layer). The dot product was then softmaxed over all the words in the passage. Likewise, the probability of a word j being the end of the answer span was computed in a similar way but using the end vector E [48]. Invalid span predictions were ignored, such as predicting an end token position that precedes a start token position, or predicting a start/end token position in the question part of the input/output sequence. Spans with top scoring probabilities were returned as a ranked list of predicted answers (or answer components) for the given question. The training objective was to minimize the sum of the softmax cross entropy loss for predicting the start and end token positions. Further details about the fine-tuning procedure are described in the context of subsection 5.2.1.

5.2. Experimental Evaluation of the Reader

In this section, we describe the setup of our experiments, then present the evaluation results (adopting the passage-scope for evaluation described in Section 3.2.3.4) and discuss them and their implications in the context of addressing the three research questions listed below in **black**. This is followed by a performance analysis of the best performing model, in which we discuss some failure and success examples to draw insight into future directions to address the identified challenges.

RQ1: Would expanding the Qur'anic passages with their corresponding Qur'an related MSA resources help the retriever in bridging the gap between the questions in MSA and their answer-bearing Qur'anic passages?

RQ2: Does further pre-training with Classical Arabic improve the performance over the MSA-only pre-trained model?

RQ3: Would it be enough to exclusively rely on transfer learning from MSA to CA in fine-tuning the readers without the need for MRC datasets in Classical Arabic?

RQ4: Adopting the passage-scope for evaluation, how does the fine-tuned CL-AraBERT reader perform on multi-answer questions vs. single-answer questions?

RQ5: Adopting the the Qur'an-scope for evaluation, how does the end-to-end QA system perform on multi-answer questions vs. single-answer questions?

RQ6: Is a native BERT-based model architecture fine-tuned as an extractive MRC reader sub-optimal for QA and MRC tasks over multi-answer questions?

5.2.1. Experimental Setup

Data Splits. We have adopted two experimental setups to perform our evaluation experiments. In the first setup, denoted as the ***holdout*** setup, we randomly split the questions (or, more-precisely, question-passage pairs) in *QRCD* into training (75%) and testing or holdout (25%) sets, as shown in Table 5.1. Adopting the passage-scope, the holdout dataset is composed of 348 question-passage-answer triplets, 227 of which are for single-answer questions and the remaining 121 are for multi-answer questions. In the second setup, denoted as the ***cross validation*** (or ***CV***) setup, we conduct a 5-fold cross validation to better evaluate the *general* performance of our model on unseen questions.

Table 5.1. Distribution of question-passage-answer triplets in *QRCD* (adopting passage-scope).

Dataset	# Question-Passage	# question-passage-answer triplets		
	Pairs	All Questions	Single-answer Questions	Multi-answer Questions
All	1093	1337	949	388
Training	819	989	722	267
Test / Holdout	274	348	227	121

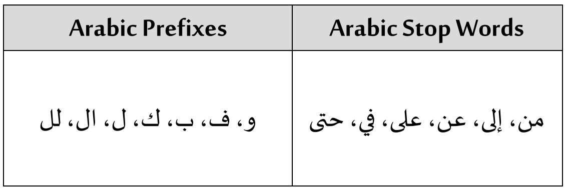
Naturally, two different random seeds were used to generate the holdout split and the CV folds. All experiments were implemented and evaluated using both setups. We elected to only release the training set of the dataset (i.e., the 75% of the entire dataset) and not the heldout set, to allow for leader-board evaluation using the holdout set, and for organizing a shared task using both, the training and holdout sets.[8]

Preprocessing. To adapt the *QRCD* dataset to the CL-AraBERT model (or any other BERT-like model), every split/fold of the dataset to be used for fine-tuning was preprocessed such that a question-passage-answer triplet was created for *each* answer span. For SQuAD v1.1, Rajpurkar, Zhang, Lopyrev, *et al.* [111] did not need to conduct this preprocessing step prior to fine-tuning/training because their dataset did not include multi-answer questions, and the answer spans for each question were variants of the same answer that may exclude/include non-essential phrases.

Evaluation Issues. To account for any relative high variation in the reported performance across folds in the CV setup, we merged the evaluation scores of the question-passage-answer triplets in each of the five test folds, before reporting their

[8] `https://sites.google.com/view/quran-qa-2022`

average over all questions in each fine-tuning experiment/run. For all fine-tuning experiments, we trained for 4 epochs using a learning rate of 3e-5 and a batch size of 32 examples. Each of the fine-tuning runs was performed five times with a different random seed for each run in both setups. Then the median performance among the five runs was reported per evaluation metric over all questions. As indicated in Section 3.2.3, Partial Average Precision (*pAP*) was the rank-based measure used for evaluating multi-answer and single-answer questions, whereas $F_1@1$ and *EM* were the set-based measures used for evaluating single-answer questions only.



Arabic Prefixes	Arabic Stop Words
و، ف، ب، ك، ل، ال، لل	من، إلى، عن، على، في، حتى

Figure 5.1. The Arabic prefixes and stopwords removed before comparing the predicted and gold answers during evaluation.

We note that before applying the partial matching procedure described in Section 3.2.3.1 during evaluation, the Farasa tool [3] was used to identify and remove prefixes from the predicted and gold answers. Removing punctuation and very common stopwords was then applied as an additional preprocessing step. This was essential to avoid mismatch due to the prefixes being included or left out from the the beginning of the gold answers during their extraction by the annotators. The prefixes and stopwords that were removed are shown in Figure 5.1.

Fine-tuning Setups. To address the above research questions, we conduct a pipelined fine-tuning procedure for both AraBERT and CL-AraBERT models using three training MRC datasets. The MSA datasets used in fine-tuning include the translated Arabic-SQuAD and the ARCD-train datasets which are composed of 48.3k and 693 question-passage-answer triplets, respectively. Overall, we have 3 different fine-tuning setups.

- fine-tuning on MSA datasets only
- fine-tuning on *QRCD* only
- fine-tuning on MSA datasets followed by further fine-tuning on *QRCD*

For ease of reference to these models, we append the term "*qrcd*", "msa" or "msa+*qrcd*" as subscripted suffixes to indicate the datasets that were used in their fine-tuning. For example, AraBERT$_{\text{msa+}qrcd}$ is the fine-tuned model using the two MSA datasets (Arabic SQuAd and ARCD) followed by the *QRCD* dataset.

5.2.2. Results and Discussion

Tables 5.2 and 5.3 present the evaluation results of the AraBERT and CL-AraBERT models over the *QRCD* dataset in the two different setups. In the subsections below, we have compared and analyzed the differences in the evaluation results after testing their statistical significance using the paired Student-t test at a confidence level of 95%.

5.2.2.1. Comparing performance of CL-AraBERT to AraBERT (RQ2)

We start by addressing *RQ2*, which is concerned with observing the effect of further pre-training the MSA pre-trained model with Classical Arabic data. Table 5.2 presents the overall performance of both models over the *QRCD* dataset in the different setups.

The results reveal two interesting observations. First, we notice that all versions of the fine-tuned classical models attained higher *pAP* scores than their counter AraBERT

Table 5.2. Results of the fine-tuned CL-AraBERT and AraBERT readers on the *QRCD* dataset. The suffixed subscripts to each model name indicate the dataset(s) used in its fine-tuning. For brevity, the subscript "msa" refers to the combined Arabic-SQuAD and ARCD datasets, and "*qrcd*" to *QRCD*. In each setup, differences between the scores annotated with the same model reference letter are statistically significant. Best results are boldfaced for each experimental setup.

Model	Fine-tuning	Holdout Setup	CV Setup
	Datasets	$pAP@10$	$pAP@10$
(a) AraBERT$_{msa}$	MSA	39.96^{cdf}	34.67^{bcdef}
(b) AraBERT$_{qrcd}$	QRCD	36.75^{cdef}	42.15^{acdef}
(c) AraBERT$_{msa+qrcd}$	MSA+QRCD	45.37^{abef}	49.53^{abdef}
(d) CL-AraBERT$_{msa}$	MSA	47.26^{abe}	39.51^{abcef}
(e) CL-AraBERT$_{qrcd}$	QRCD	40.66^{bcdf}	44.88^{abcdf}
(f) CL-AraBERT$_{msa+qrcd}$	MSA+QRCD	$\mathbf{51.49}^{abce}$	$\mathbf{53.28}^{abcde}$

models that were fine-tuned in the same way. The differences between these scores were all statistically significant. For example, CL-AraBERT$_{\text{msa}}$ attained a lead of 7.3 and 4.8 points on its pAP scores over AraBERT$_{\text{msa}}$ in the holdout and CV setups, respectively (Table 5.2). This finding suggests that the classical model consistently outperforms the other non-classical one on the *QRCD* dataset when both models undergo the same fine-tuning procedure. As such, we can affirm that such improvements in performance are mainly attributed to the further classical pre-training using a large segment from the Classical Arabic corpus OpenITI [114].

Second, among all models, CL-AraBERT$_{\text{msa+}qrcd}$ attained the best pAP scores in the two experimental setups, achieving an improvement of 6.1 and 3.8 points over AraBERT$_{\text{msa+}qrcd}$ in the hold-out and the CV setups, respectively. This shows the importance of fine-tuning using *both* non-classical and classical MRC training sets along side the classical pre-training. We address this further in the next section.

5.2.2.2. Transfer learning from MSA to Classical Arabic (RQ3)

We address the third research question (RQ3), that is concerned with observing the gains in performance due to cross-lingual transfer learning, by comparing the performance of the pre-trained models that are fine-tuned using both *QRCD* and MSA datasets with the models that are fine-tuned using only one of them.

We start by comparing the performance of AraBERT$_{qrcd}$ reader to the AraBERT$_{\text{msa}+qrcd}$ reader. The latter model attained better pAP scores than the former by 8.6 and 7.4 points in the holdout and CV setups, respectively (Table 5.2). Similar improvements were also witnessed by CL-AraBERT$_{\text{msa}+qrcd}$ in comparison to CL-AraBERT$_{qrcd}$ as shown in Table 5.2. These statistically significant differences over the pAP evaluation scores are considered gains in performance, which were conquered due to fine-tuning using the relatively large reading comprehension MSA dataset. The Arabic SQuAD dataset provided 48.3k question-passage-answer triplets, while the ARCD-train dataset provided another 693 triplets as training examples [97].

However, relying exclusively on MRC datasets in MSA only (without MRC datasets in Classical Arabic) may not be sufficient for our MRC task on the Holy Qur'an. Comparing the performance of AraBERT$_{\text{msa}+qrcd}$ and CL-AraBERT$_{\text{msa}+qrcd}$ with their counter models that were exclusively fine-tuned using the two MSA datasets, has revealed this gap, especially in the CV setup. AraBERT$_{\text{msa}+qrcd}$ outperformed AraBERT$_{\text{msa}}$ by $\sim$14.9 points on its pAP score (Table 5.2). Likewise, CL-AraBERT$_{\text{msa}+qrcd}$ outperformed CL-AraBERT$_{\text{msa}}$ by $\sim$13.8 points on its pAP score (Table 5.2). The pAP scores in the holdout setup have also revealed this difference in performance, but with a lesser extent.

While the performance using MSA-only datasets is fair, the above findings demonstrate the impact of the *QRCD* dataset (as a Classical Arabic resource) in boosting performance of classical and non-classical models, despite its relatively modest size of 1,337 question-passage-answer triplets. They also suggest that MSA resources can be used in transfer learning to enhance the performance of MRC tasks on the Holy Qur'an, but it would be essential to complement them with Classical Arabic resources as well to attain better performance. Any gains due to transfer learning could be mainly attributed to the existing similarity between MSA and Classical Arabic with respect to morphology and syntax characteristics. Nevertheless, Classical Arabic remains richer in lexis [121], despite the contemporary western words that found their way into MSA through translation or transliteration.

5.2.2.3. Performance across question types (RQ4)

With 14% of the question-passage pairs in *QRCD* comprising two or more answers (according to the passage-scope), it was imperative to address our fourth research question regarding the performance of CL-AraBERT over multi-answer questions in comparison to single-answer questions.

Table 5.3 presents the comparison in terms of all possible measures over both experimental setups. It is clearly noted that, in both setups, all the fine-tuned models performed better, in terms of pAP, on single-answer questions in comparison to multi-answer questions. This is not very surprising given that the majority of the training examples in *QRCD* and all the training examples in the two MSA datasets are for single-

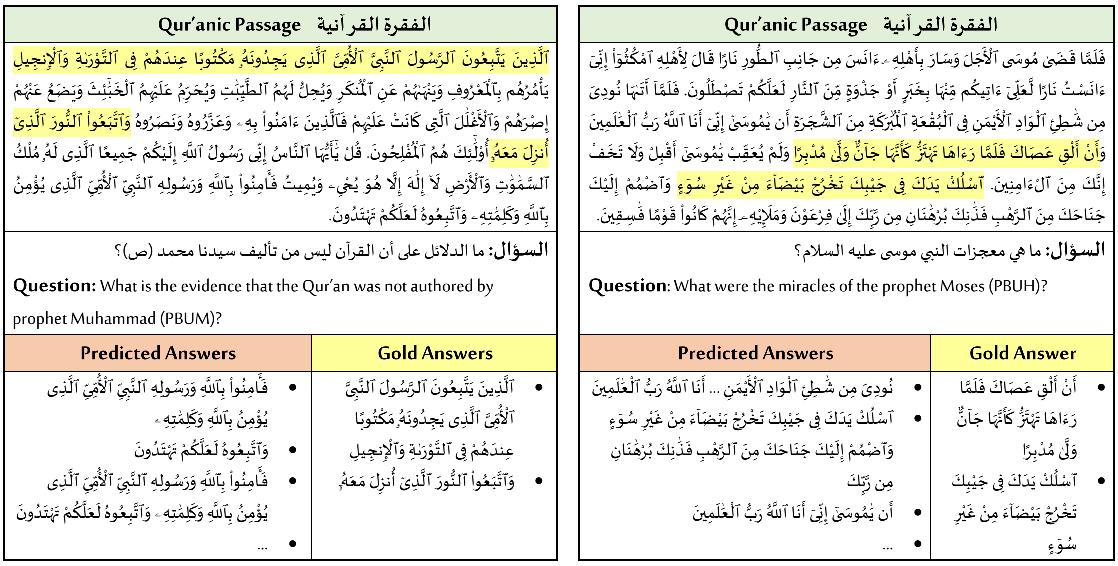
Table 5.3. Results of the fine-tuned CL-AraBERT and AraBERT readers across question types in the *QRCD* dataset. The letters "S" and "M" correspond to "single-answer" and "multi-answer" questions, respectively. In each column, differences between the scores annotated with the same model reference letter are statistically significant. Best results are boldfaced in each experimental setup.

Model	Qst.	Holdout Setup			Cross-Validation Setup		
	Type	F_1@1	*EM*	pAP@10	F_1@1	*EM*	pAP@10
(a) AraBERT$_{msa}$	S	38.72^{f}	11.50^{cf}	41.90^{cdf}	32.59^{bcdef}	10.22^{bcdef}	35.41^{bcdef}
	M			27.50^{df}			30.16^{bcdef}
(b) AraBERT$_{qrcd}$	S	30.89^{cdf}	11.50^{cdf}	37.77^{cdf}	37.55^{acef}	19.28^{acdf}	42.74^{acef}
	M			31.96^{f}			37.3^{acf}
(c) AraBERT$_{msa+qrcd}$	S	41.99^{be}	18.14^{ab}	47.45^{abe}	45.84^{abde}	26.84^{abde}	50.42^{abdef}
	M			37.66			45.01^{abde}
(d) CL-AraBERT$_{msa}$	S	45.68^{be}	19.03^{b}	48.97^{abe}	36.98^{acef}	14.59^{abcef}	40.18^{acef}
	M			37.47^{af}			34.56^{acf}
(e) CL-AraBERT$_{qrcd}$	S	34.85^{cdf}	15.49^{f}	41.40^{cdf}	40.94^{abcdf}	21.19^{acdf}	45.61^{abcdf}
	M			35.76^{f}			40.25^{acf}
(f) CL-AraBERT$_{msa+qrcd}$	S	$\mathbf{47.25}^{abe}$	$\mathbf{23.89}^{abe}$	$\mathbf{52.44}^{abce}$	$\mathbf{49.68}^{abde}$	$\mathbf{28.01}^{abde}$	$\mathbf{53.97}^{abcde}$
	M			$\mathbf{47.53}^{abde}$			$\mathbf{47.40}^{abde}$

answer questions. Moreover, multi-answer questions are naturally more challenging, hence typically harder. Again, CL-AraBERT$_{\text{msa+}qrcd}$ was the pioneer in outperforming all the other models on both question types by attaining the highest pAP, F_1@1 and *EM* scores. Its pAP scores on single-answer questions were better than those on multi-answer questions by 4.9 points in the holdout setup, and 6.6 points in the CV setup.

In general, we note that the range of the *EM* scores, in comparison to the $F_1@1$ and pAP scores in Table 5.3, was the lowest (ranging from 10.22 to 28.01 points), while the range of the $F_1@1$ scores was relatively higher (ranging from 30.89 to 49.68). This makes the range of the pAP scores the highest (ranging from 27.50 53.97). This finding suggests that the pAP evaluation measure could be the most sensitive to improvement/deterioration in performance because it is rank-based and inherently sensitive to partial/exact matches, which in turn makes it less stringent than the *EM* and $F_1@1$ set-based measures. The latter two measures are considered stringent because they only consider the top prediction in the evaluation, with $F_1@1$ more lenient as it rewards partial matching.

5.2.2.4. Performance Analysis of the Reader



Qur’anic Passage الفقرة القرآنية

ٱلَّذِينَ يَتَّبِعُونَ ٱلرَّسُولَ ٱلنَّبِيَّ ٱلْأُمِّيَّ ٱلَّذِى يَجِدُونَهُۥ مَكْتُوبًا عِندَهُمْ فِى ٱلتَّوْرَىٰةِ وَٱلْإِنجِيلِ يَأْمُرُهُم بِٱلْمَعْرُوفِ وَيَنْهَىٰهُمْ عَنِ ٱلْمُنكَرِ وَيُحِلُّ لَهُمُ ٱلطَّيِّبَٰتِ وَيُحَرِّمُ عَلَيْهِمُ ٱلْخَبَٰٓئِثَ وَيَضَعُ عَنْهُمْ إِصْرَهُمْ وَٱلْأَغْلَٰلَ ٱلَّتِى كَانَتْ عَلَيْهِمْ فَٱلَّذِينَ ءَامَنُوا۟ بِهِۦ وَعَزَّرُوهُ وَنَصَرُوهُ وَٱتَّبَعُوا۟ ٱلنُّورَ ٱلَّذِىٓ أُنزِلَ مَعَهُۥٓ أُو۟لَٰٓئِكَ هُمُ ٱلْمُفْلِحُونَ. قُلْ يَٰٓأَيُّهَا ٱلنَّاسُ إِنِّى رَسُولُ ٱللَّهِ إِلَيْكُمْ جَمِيعًا ٱلَّذِى لَهُۥ مُلْكُ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ لَآ إِلَٰهَ إِلَّا هُوَ يُحْىِۦ وَيُمِيتُ فَـَٔامِنُوا۟ بِٱللَّهِ وَرَسُولِهِ ٱلنَّبِىِّ ٱلْأُمِّىِّ ٱلَّذِى يُؤْمِنُ بِٱللَّهِ وَكَلِمَٰتِهِۦ وَٱتَّبِعُوهُ لَعَلَّكُمْ تَهْتَدُونَ.

السؤال: ما الدلائل على أن القرآن ليس من تأليف سيدنا محمد (ص)؟

Question: What is the evidence that the Qur’an was not authored by prophet Muhammad (PBUM)?

Predicted Answers	Gold Answers
• فَـَٔامِنُوا۟ بِٱللَّهِ وَرَسُولِهِ ٱلنَّبِىِّ ٱلْأُمِّىِّ ٱلَّذِى يُؤْمِنُ بِٱللَّهِ وَكَلِمَٰتِهِۦ • وَٱتَّبِعُوهُ لَعَلَّكُمْ تَهْتَدُونَ • فَـَٔامِنُوا۟ بِٱللَّهِ وَرَسُولِهِ ٱلنَّبِىِّ ٱلْأُمِّىِّ ٱلَّذِى يُؤْمِنُ بِٱللَّهِ وَكَلِمَٰتِهِۦ وَٱتَّبِعُوهُ لَعَلَّكُمْ تَهْتَدُونَ • ...	• ٱلَّذِينَ يَتَّبِعُونَ ٱلرَّسُولَ ٱلنَّبِيَّ ٱلْأُمِّيَّ ٱلَّذِى يَجِدُونَهُۥ مَكْتُوبًا عِندَهُمْ فِى ٱلتَّوْرَىٰةِ وَٱلْإِنجِيلِ • وَٱتَّبَعُوا۟ ٱلنُّورَ ٱلَّذِىٓ أُنزِلَ مَعَهُۥٓ

(a)

Qur’anic Passage الفقرة القرآنية

فَلَمَّا قَضَىٰ مُوسَى ٱلْأَجَلَ وَسَارَ بِأَهْلِهِۦٓ ءَانَسَ مِن جَانِبِ ٱلطُّورِ نَارًا قَالَ لِأَهْلِهِ ٱمْكُثُوٓا۟ إِنِّىٓ ءَانَسْتُ نَارًا لَّعَلِّىٓ ءَاتِيكُم مِّنْهَا بِخَبَرٍ أَوْ جَذْوَةٍ مِّنَ ٱلنَّارِ لَعَلَّكُمْ تَصْطَلُونَ. فَلَمَّآ أَتَىٰهَا نُودِىَ مِن شَٰطِئِ ٱلْوَادِ ٱلْأَيْمَنِ فِى ٱلْبُقْعَةِ ٱلْمُبَٰرَكَةِ مِنَ ٱلشَّجَرَةِ أَن يَٰمُوسَىٰٓ إِنِّىٓ أَنَا ٱللَّهُ رَبُّ ٱلْعَٰلَمِينَ وَأَنْ أَلْقِ عَصَاكَ فَلَمَّا رَءَاهَا تَهْتَزُّ كَأَنَّهَا جَآنٌّ وَلَّىٰ مُدْبِرًا وَلَمْ يُعَقِّبْ يَٰمُوسَىٰٓ أَقْبِلْ وَلَا تَخَفْ إِنَّكَ مِنَ ٱلْءَامِنِينَ. ٱسْلُكْ يَدَكَ فِى جَيْبِكَ تَخْرُجْ بَيْضَآءَ مِنْ غَيْرِ سُوٓءٍ وَٱضْمُمْ إِلَيْكَ جَنَاحَكَ مِنَ ٱلرَّهْبِ فَذَٰنِكَ بُرْهَٰنَانِ مِن رَّبِّكَ إِلَىٰ فِرْعَوْنَ وَمَلَإِي۟هِۦٓ إِنَّهُمْ كَانُوا۟ قَوْمًا فَٰسِقِينَ.

السؤال: ما هي معجزات النبي موسى عليه السلام؟

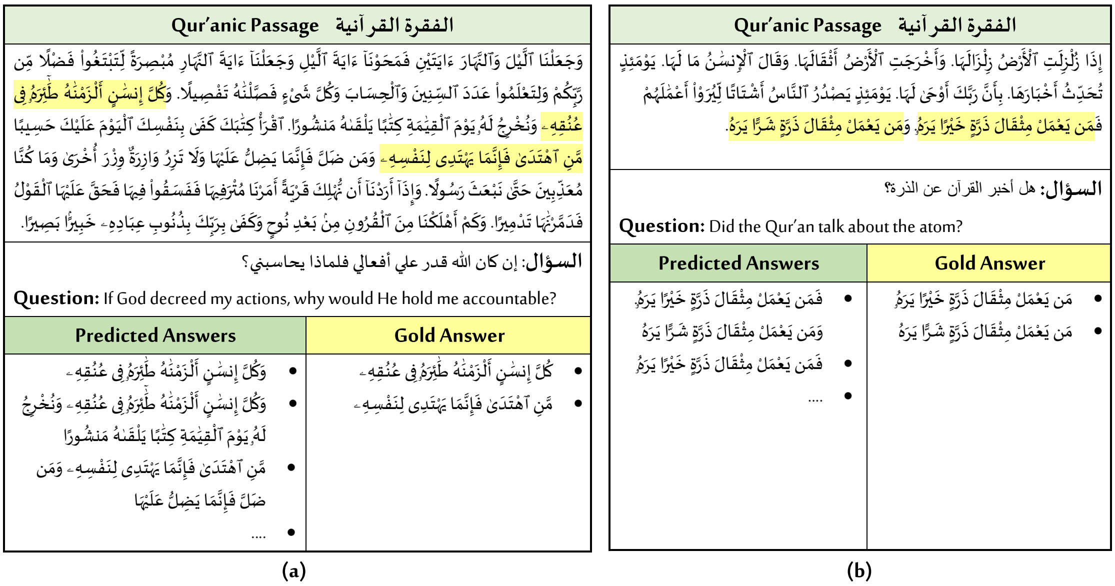
Question: What were the miracles of the prophet Moses (PBUH)?

Predicted Answers	Gold Answer
• نُودِىَ مِن شَٰطِئِ ٱلْوَادِ ٱلْأَيْمَنِ ... أَنَا ٱللَّهُ رَبُّ ٱلْعَٰلَمِينَ • ٱسْلُكْ يَدَكَ فِى جَيْبِكَ تَخْرُجْ بَيْضَآءَ مِنْ غَيْرِ سُوٓءٍ وَٱضْمُمْ إِلَيْكَ جَنَاحَكَ مِنَ ٱلرَّهْبِ فَذَٰنِكَ بُرْهَٰنَانِ مِن رَّبِّكَ • أَن يَٰمُوسَىٰٓ إِنِّىٓ أَنَا ٱللَّهُ رَبُّ ٱلْعَٰلَمِينَ • ...	• أَنْ أَلْقِ عَصَاكَ فَلَمَّا رَءَاهَا تَهْتَزُّ كَأَنَّهَا جَآنٌّ وَلَّىٰ مُدْبِرًا • ٱسْلُكْ يَدَكَ فِى جَيْبِكَ تَخْرُجْ بَيْضَآءَ مِنْ غَيْرِ سُوٓءٍ

(b)

Figure 5.2. A failure example (a) and a semi-failure example (b) of multi-answer questions. The first was incorrectly answered and the second was partially answered by CL-AraBERT$_{\text{msa+}qrcd}$.

In this section, we discuss and present several failure and success examples (in Figures 5.2 through 5.5) in an attempt to understand the weaknesses and strengths of



Qur'anic Passage الفقرة القرآنية

وَجَعَلْنَا ٱلَّيْلَ وَٱلنَّهَارَ ءَايَتَيْنِ فَمَحَوْنَآ ءَايَةَ ٱلَّيْلِ وَجَعَلْنَآ ءَايَةَ ٱلنَّهَارِ مُبْصِرَةً لِّتَبْتَغُوا۟ فَضْلًا مِّن رَّبِّكُمْ وَلِتَعْلَمُوا۟ عَدَدَ ٱلسِّنِينَ وَٱلْحِسَابَ وَكُلَّ شَىْءٍ فَصَّلْنَٰهُ تَفْصِيلًا. وَكُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ وَنُخْرِجُ لَهُۥ يَوْمَ ٱلْقِيَٰمَةِ كِتَٰبًا يَلْقَىٰهُ مَنشُورًا. ٱقْرَأْ كِتَٰبَكَ كَفَىٰ بِنَفْسِكَ ٱلْيَوْمَ عَلَيْكَ حَسِيبًا مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا وَلَا تَزِرُ وَازِرَةٌ وِزْرَ أُخْرَىٰ وَمَا كُنَّا مُعَذِّبِينَ حَتَّىٰ نَبْعَثَ رَسُولًا. وَإِذَآ أَرَدْنَآ أَن نُّهْلِكَ قَرْيَةً أَمَرْنَا مُتْرَفِيهَا فَفَسَقُوا۟ فِيهَا فَحَقَّ عَلَيْهَا ٱلْقَوْلُ فَدَمَّرْنَٰهَا تَدْمِيرًا. وَكَمْ أَهْلَكْنَا مِنَ ٱلْقُرُونِ مِنۢ بَعْدِ نُوحٍ وَكَفَىٰ بِرَبِّكَ بِذُنُوبِ عِبَادِهِۦ خَبِيرًۢا بَصِيرًا.

السؤال: إن كان الله قدر علي أفعالي فلماذا يحاسبني؟

Question: If God decreed my actions, why would He hold me accountable?

Predicted Answers	Gold Answer
• وَكُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ • وَكُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ وَنُخْرِجُ لَهُۥ يَوْمَ ٱلْقِيَٰمَةِ كِتَٰبًا يَلْقَىٰهُ مَنشُورًا • مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا •	• كُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ • مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ

(a)

Qur'anic Passage الفقرة القرآنية

إِذَا زُلْزِلَتِ ٱلْأَرْضُ زِلْزَالَهَا. وَأَخْرَجَتِ ٱلْأَرْضُ أَثْقَالَهَا. وَقَالَ ٱلْإِنسَٰنُ مَا لَهَا. يَوْمَئِذٍ تُحَدِّثُ أَخْبَارَهَا. بِأَنَّ رَبَّكَ أَوْحَىٰ لَهَا. يَوْمَئِذٍ يَصْدُرُ ٱلنَّاسُ أَشْتَاتًا لِّيُرَوْا۟ أَعْمَٰلَهُمْ فَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ خَيْرًا يَرَهُۥ. وَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ شَرًّا يَرَهُۥ.

السؤال: هل أخبر القرآن عن الذرة؟

Question: Did the Qur'an talk about the atom?

Predicted Answers	Gold Answer
• فَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ خَيْرًا يَرَهُۥ. وَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ شَرًّا يَرَهُۥ • فَمَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ خَيْرًا يَرَهُۥ •	• مَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ خَيْرًا يَرَهُۥ • مَن يَعْمَلْ مِثْقَالَ ذَرَّةٍ شَرًّا يَرَهُۥ

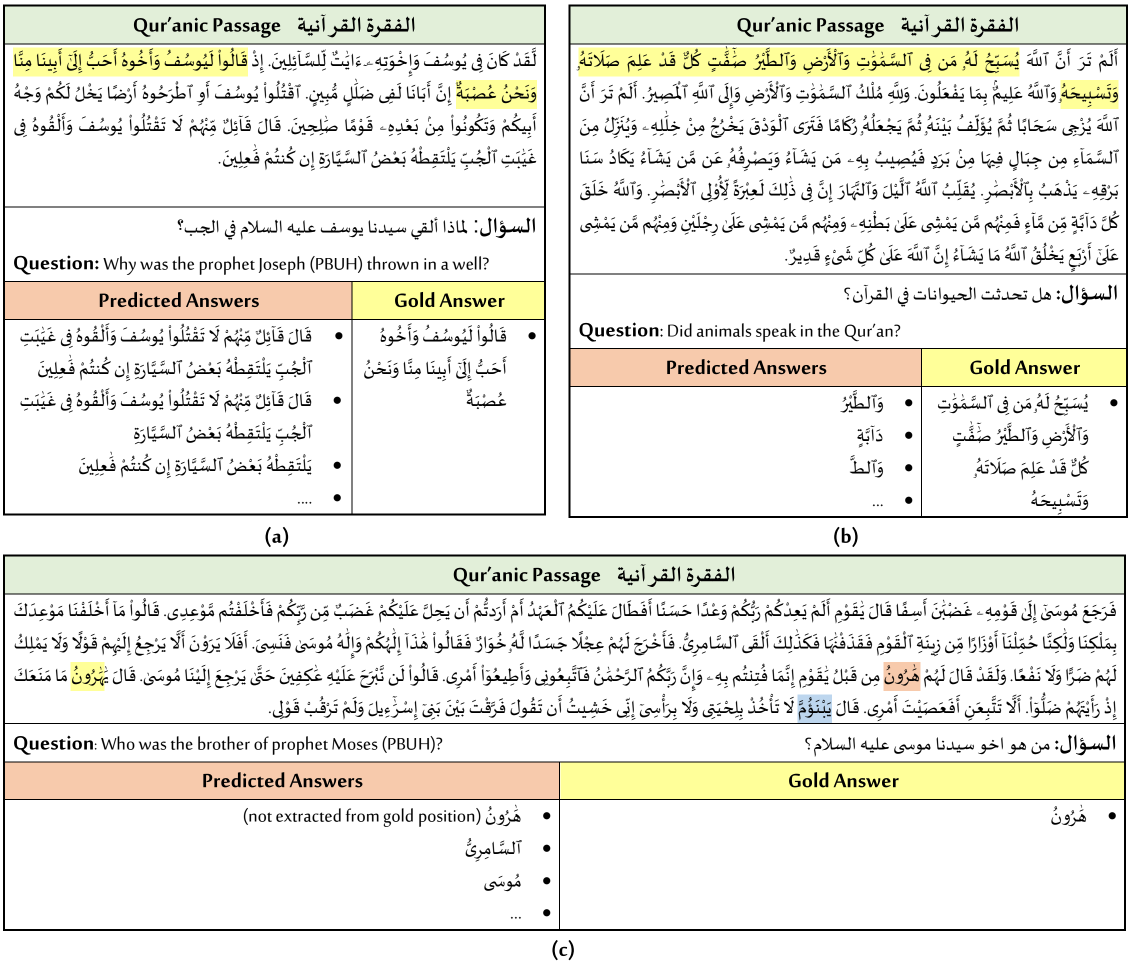
(b)

Figure 5.3. Two success examples of multi-answer questions correctly answered by CL-AraBERT$_{\text{msa}+qrcd}$.

the fine-tuned CL-AraBERT$_{\text{msa}+qrcd}$ reader model (since it is the best performing model) on the *QRCD* dataset. This performance analysis would provide insights towards future directions to build on its strengths and address its weaknesses.

We recall that multi-answer and single-answer questions in *QRCD* comprise factoid and non-factoid question types that include list, causal, definition, yes/no questions, and beyond. Failure to answer some questions could be attributed to one or more of the following challenges, though CL-AraBERT$_{\text{msa}+qrcd}$ was able to overcome some of these challenges for other questions, as demonstrated in the success examples:

(1) **Evidence-based answers**. While the literary style of the Qur'anic verses may resonate very well with the answer types of factoid questions, they may not fully comply with traditional natural language answers to non-factoid questions. This would tend to make answering such questions more challenging. For example, answer(s) to a yes/no question can only be drawn from Qur'anic verses that provide evidence that asserts or negates that question. In general, answers



Qur'anic Passage الفقرة القرآنية	
لَّقَدْ كَانَ فِى يُوسُفَ وَإِخْوَتِهِۦٓ ءَايَٰتٌ لِّلسَّآئِلِينَ. إِذْ قَالُوا۟ لَيُوسُفُ وَأَخُوهُ أَحَبُّ إِلَىٰٓ أَبِينَا مِنَّا وَنَحْنُ عُصْبَةٌ إِنَّ أَبَانَا لَفِى ضَلَٰلٍ مُّبِينٍ. ٱقْتُلُوا۟ يُوسُفَ أَوِ ٱطْرَحُوهُ أَرْضًا يَخْلُ لَكُمْ وَجْهُ أَبِيكُمْ وَتَكُونُوا۟ مِنۢ بَعْدِهِۦ قَوْمًا صَٰلِحِينَ. قَالَ قَآئِلٌ مِّنْهُمْ لَا تَقْتُلُوا۟ يُوسُفَ وَأَلْقُوهُ فِى غَيَٰبَتِ ٱلْجُبِّ يَلْتَقِطْهُ بَعْضُ ٱلسَّيَّارَةِ إِن كُنتُمْ فَٰعِلِينَ.	
السؤال: لماذا ألقي سيدنا يوسف عليه السلام في الجب؟ **Question:** Why was the prophet Joseph (PBUH) thrown in a well?	
Predicted Answers	**Gold Answer**
• قَالَ قَآئِلٌ مِّنْهُمْ لَا تَقْتُلُوا۟ يُوسُفَ وَأَلْقُوهُ فِى غَيَٰبَتِ ٱلْجُبِّ يَلْتَقِطْهُ بَعْضُ ٱلسَّيَّارَةِ إِن كُنتُمْ فَٰعِلِينَ • قَالَ قَآئِلٌ مِّنْهُمْ لَا تَقْتُلُوا۟ يُوسُفَ وَأَلْقُوهُ فِى غَيَٰبَتِ ٱلْجُبِّ يَلْتَقِطْهُ بَعْضُ ٱلسَّيَّارَةِ • يَلْتَقِطْهُ بَعْضُ ٱلسَّيَّارَةِ إِن كُنتُمْ فَٰعِلِينَ •	• قَالُوا۟ لَيُوسُفُ وَأَخُوهُ أَحَبُّ إِلَىٰٓ أَبِينَا مِنَّا وَنَحْنُ عُصْبَةٌ

(a)

Qur'anic Passage الفقرة القرآنية	
أَلَمْ تَرَ أَنَّ ٱللَّهَ يُسَبِّحُ لَهُۥ مَن فِى ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ وَٱلطَّيْرُ صَٰٓفَّٰتٍ كُلٌّ قَدْ عَلِمَ صَلَاتَهُۥ وَتَسْبِيحَهُۥ وَٱللَّهُ عَلِيمٌۢ بِمَا يَفْعَلُونَ. وَلِلَّهِ مُلْكُ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ وَإِلَى ٱللَّهِ ٱلْمَصِيرُ. أَلَمْ تَرَ أَنَّ ٱللَّهَ يُزْجِى سَحَابًا ثُمَّ يُؤَلِّفُ بَيْنَهُۥ ثُمَّ يَجْعَلُهُۥ رُكَامًا فَتَرَى ٱلْوَدْقَ يَخْرُجُ مِنْ خِلَٰلِهِۦ وَيُنَزِّلُ مِنَ ٱلسَّمَآءِ مِن جِبَالٍ فِيهَا مِنۢ بَرَدٍ فَيُصِيبُ بِهِۦ مَن يَشَآءُ وَيَصْرِفُهُۥ عَن مَّن يَشَآءُ يَكَادُ سَنَا بَرْقِهِۦ يَذْهَبُ بِٱلْأَبْصَٰرِ. يُقَلِّبُ ٱللَّهُ ٱلَّيْلَ وَٱلنَّهَارَ إِنَّ فِى ذَٰلِكَ لَعِبْرَةً لِّأُو۟لِى ٱلْأَبْصَٰرِ. وَٱللَّهُ خَلَقَ كُلَّ دَآبَّةٍ مِّن مَّآءٍ فَمِنْهُم مَّن يَمْشِى عَلَىٰ بَطْنِهِۦ وَمِنْهُم مَّن يَمْشِى عَلَىٰ رِجْلَيْنِ وَمِنْهُم مَّن يَمْشِى عَلَىٰٓ أَرْبَعٍ يَخْلُقُ ٱللَّهُ مَا يَشَآءُ إِنَّ ٱللَّهَ عَلَىٰ كُلِّ شَىْءٍ قَدِيرٌ.	
السؤال: هل تحدثت الحيوانات في القرآن؟ **Question**: Did animals speak in the Qur'an?	
Predicted Answers	**Gold Answer**
• وَٱلطَّيْرُ • دَآبَّةٍ • وَٱلطَّ • ...	• يُسَبِّحُ لَهُۥ مَن فِى ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ وَٱلطَّيْرُ صَٰٓفَّٰتٍ كُلٌّ قَدْ عَلِمَ صَلَاتَهُۥ وَتَسْبِيحَهُۥ

(b)

Qur'anic Passage الفقرة القرآنية	
فَرَجَعَ مُوسَىٰٓ إِلَىٰ قَوْمِهِۦ غَضْبَٰنَ أَسِفًا قَالَ يَٰقَوْمِ أَلَمْ يَعِدْكُمْ رَبُّكُمْ وَعْدًا حَسَنًا أَفَطَالَ عَلَيْكُمُ ٱلْعَهْدُ أَمْ أَرَدتُّمْ أَن يَحِلَّ عَلَيْكُمْ غَضَبٌ مِّن رَّبِّكُمْ فَأَخْلَفْتُم مَّوْعِدِى. قَالُوا۟ مَآ أَخْلَفْنَا مَوْعِدَكَ بِمَلْكِنَا وَلَٰكِنَّا حُمِّلْنَآ أَوْزَارًا مِّن زِينَةِ ٱلْقَوْمِ فَقَذَفْنَٰهَا فَكَذَٰلِكَ أَلْقَى ٱلسَّامِرِىُّ. فَأَخْرَجَ لَهُمْ عِجْلًا جَسَدًا لَّهُۥ خُوَارٌ فَقَالُوا۟ هَٰذَآ إِلَٰهُكُمْ وَإِلَٰهُ مُوسَىٰ فَنَسِىَ. أَفَلَا يَرَوْنَ أَلَّا يَرْجِعُ إِلَيْهِمْ قَوْلًا وَلَا يَمْلِكُ لَهُمْ ضَرًّا وَلَا نَفْعًا. وَلَقَدْ قَالَ لَهُمْ هَٰرُونُ مِن قَبْلُ يَٰقَوْمِ إِنَّمَا فُتِنتُم بِهِۦ وَإِنَّ رَبَّكُمُ ٱلرَّحْمَٰنُ فَٱتَّبِعُونِى وَأَطِيعُوٓا۟ أَمْرِى. قَالُوا۟ لَن نَّبْرَحَ عَلَيْهِ عَٰكِفِينَ حَتَّىٰ يَرْجِعَ إِلَيْنَا مُوسَىٰ. قَالَ يَٰهَٰرُونُ مَا مَنَعَكَ إِذْ رَأَيْتَهُمْ ضَلُّوٓا۟. أَلَّا تَتَّبِعَنِ أَفَعَصَيْتَ أَمْرِى. قَالَ يَبْنَؤُمَّ لَا تَأْخُذْ بِلِحْيَتِى وَلَا بِرَأْسِىٓ إِنِّى خَشِيتُ أَن تَقُولَ فَرَّقْتَ بَيْنَ بَنِىٓ إِسْرَٰٓءِيلَ وَلَمْ تَرْقُبْ قَوْلِى.	
Question: Who was the brother of prophet Moses (PBUH)? السؤال: من هو اخو سيدنا موسى عليه السلام؟	
Predicted Answers	**Gold Answer**
• هَٰرُونُ (not extracted from gold position) • ٱلسَّامِرِىُّ • مُوسَىٰ • ...	• هَٰرُونُ

(c)

Figure 5.4. Three failure examples of single-answer questions that were not correctly answered by CL-AraBERT$_{\text{msa+}qrcd}$. Text highlighted in blue is the reference expression to the preceding antecedent highlighted in yellow.

to non-factoid questions are mostly evidence-based in the Holy Qur'an. For the multi-answer question in Figure 5.2(a), the reader failed to return the two answers which provide evidence that prophet Muhammad (PBUH) did not author the Qur'an, while in Figure 5.3(a), it succeeded in returning the two evidence-based answers to the challenging *why* question. Another failure example and another success example related to this challenge are exhibited in Figure 5.4(b) and Figure 5.3(b), respectively.

We note that some of the examples mentioned above (such as Figure 5.4(b) and Figure 5.3(a)) may also demonstrate one or more of the challenges described in

Qur'anic Passage الفقرة القرآنية

وَرَٰوَدَتْهُ ٱلَّتِى هُوَ فِى بَيْتِهَا عَن نَّفْسِهِۦ وَغَلَّقَتِ ٱلْأَبْوَٰبَ وَقَالَتْ هَيْتَ لَكَ قَالَ مَعَاذَ ٱللَّهِ إِنَّهُۥ رَبِّىٓ أَحْسَنَ مَثْوَاىَ إِنَّهُۥ لَا يُفْلِحُ ٱلظَّٰلِمُونَ. وَلَقَدْ هَمَّتْ بِهِۦ وَهَمَّ بِهَا لَوْلَآ أَن رَّءَا بُرْهَٰنَ رَبِّهِۦ كَذَٰلِكَ لِنَصْرِفَ عَنْهُ ٱلسُّوٓءَ وَٱلْفَحْشَآءَ إِنَّهُۥ مِنْ عِبَادِنَا ٱلْمُخْلَصِينَ. وَٱسْتَبَقَا ٱلْبَابَ وَقَدَّتْ قَمِيصَهُۥ مِن دُبُرٍ وَأَلْفَيَا سَيِّدَهَا لَدَا ٱلْبَابِ قَالَتْ مَا جَزَآءُ مَنْ أَرَادَ بِأَهْلِكَ سُوٓءًا إِلَّآ أَن يُسْجَنَ أَوْ عَذَابٌ أَلِيمٌ. قَالَ هِىَ رَٰوَدَتْنِى عَن نَّفْسِى وَشَهِدَ شَاهِدٌ مِّنْ أَهْلِهَآ إِن كَانَ قَمِيصُهُۥ قُدَّ مِن قُبُلٍ فَصَدَقَتْ وَهُوَ مِنَ ٱلْكَٰذِبِينَ. وَإِن كَانَ قَمِيصُهُۥ قُدَّ مِن دُبُرٍ فَكَذَبَتْ وَهُوَ مِنَ ٱلصَّٰدِقِينَ. فَلَمَّا رَءَا قَمِيصَهُۥ قُدَّ مِن دُبُرٍ قَالَ إِنَّهُۥ مِن كَيْدِكُنَّ إِنَّ كَيْدَكُنَّ عَظِيمٌ. يُوسُفُ أَعْرِضْ عَنْ هَٰذَا وَٱسْتَغْفِرِى لِذَنۢبِكِ إِنَّكِ كُنتِ مِنَ ٱلْخَاطِـِٔينَ. وَقَالَ نِسْوَةٌ فِى ٱلْمَدِينَةِ ٱمْرَأَتُ ٱلْعَزِيزِ تُرَٰوِدُ فَتَىٰهَا عَن نَّفْسِهِۦ قَدْ شَغَفَهَا حُبًّا إِنَّا لَنَرَىٰهَا فِى ضَلَٰلٍ مُّبِينٍ. فَلَمَّا سَمِعَتْ بِمَكْرِهِنَّ أَرْسَلَتْ إِلَيْهِنَّ وَأَعْتَدَتْ لَهُنَّ مُتَّكَـًٔا وَءَاتَتْ كُلَّ وَٰحِدَةٍ مِّنْهُنَّ سِكِّينًا وَقَالَتِ ٱخْرُجْ عَلَيْهِنَّ فَلَمَّا رَأَيْنَهُۥٓ أَكْبَرْنَهُۥ وَقَطَّعْنَ أَيْدِيَهُنَّ وَقُلْنَ حَٰشَ لِلَّهِ مَا هَٰذَا بَشَرًا إِنْ هَٰذَآ إِلَّا مَلَكٌ كَرِيمٌ. قَالَتْ فَذَٰلِكُنَّ ٱلَّذِى لُمْتُنَّنِى فِيهِ وَلَقَدْ رَٰوَدتُّهُۥ عَن نَّفْسِهِۦ فَٱسْتَعْصَمَ وَلَئِن لَّمْ يَفْعَلْ مَآ ءَامُرُهُۥ لَيُسْجَنَنَّ وَلَيَكُونًا مِّنَ ٱلصَّٰغِرِينَ. قَالَ رَبِّ ٱلسِّجْنُ أَحَبُّ إِلَىَّ مِمَّا يَدْعُونَنِىٓ إِلَيْهِ وَإِلَّا تَصْرِفْ عَنِّى كَيْدَهُنَّ أَصْبُ إِلَيْهِنَّ وَأَكُن مِّنَ ٱلْجَٰهِلِينَ.

السؤال: من هو النبي الذى دخل السجن؟

Question: Who was the prophet that went to prison?

Predicted Answers	Gold Answer
• يُوسُفُ • يُوسُفُ . • يُوسُفُ أَعْرِضْ عَنْ هَٰذَا وَٱسْتَغْفِرِى لِذَنۢبِكِ إِنَّكِ كُنتِ مِنَ ٱلْخَاطِـِٔينَ •	• يُوسُفُ

(a)

Qur'anic Passage الفقرة القرآنية

أَرَءَيْتَ ٱلَّذِى يَنْهَىٰ. عَبْدًا إِذَا صَلَّىٰٓ. أَرَءَيْتَ إِن كَانَ عَلَى ٱلْهُدَىٰٓ. أَوْ أَمَرَ بِٱلتَّقْوَىٰٓ. أَرَءَيْتَ إِن كَذَّبَ وَتَوَلَّىٰٓ. أَلَمْ يَعْلَم بِأَنَّ ٱللَّهَ يَرَىٰ. كَلَّا لَئِن لَّمْ يَنتَهِ لَنَسْفَعًۢا بِٱلنَّاصِيَةِ. نَاصِيَةٍ كَٰذِبَةٍ خَاطِئَةٍ. فَلْيَدْعُ نَادِيَهُۥ. سَنَدْعُ ٱلزَّبَانِيَةَ. كَلَّا لَا تُطِعْهُ وَٱسْجُدْ وَٱقْتَرِب.

السؤال/Question: ما هي الإشارات للدماغ أو لأجزاء من الدماغ في القرآن؟

Q: What are the references to the brain or parts of the brain in the Qur'an?

Predicted Answers	Gold Answer
• بِٱلنَّاصِيَةِ • نَاصِيَةٍ كَٰذِبَةٍ خَاطِئَةٍ • ناصية •	• ٱلنَّاصِيَةِ (أو نَاصِيَةٍ)

(b)

Qur'anic Passage الفقرة القرآنية

وَٱذْكُرْ عَبْدَنَآ أَيُّوبَ إِذْ نَادَىٰ رَبَّهُۥٓ أَنِّى مَسَّنِىَ ٱلشَّيْطَٰنُ بِنُصْبٍ وَعَذَابٍ. ٱرْكُضْ بِرِجْلِكَ هَٰذَا مُغْتَسَلٌۢ بَارِدٌ وَشَرَابٌ. وَوَهَبْنَا لَهُۥٓ أَهْلَهُۥ وَمِثْلَهُم مَّعَهُمْ رَحْمَةً مِّنَّا وَذِكْرَىٰ لِأُو۟لِى ٱلْأَلْبَٰبِ. وَخُذْ بِيَدِكَ ضِغْثًا فَٱضْرِب بِّهِۦ وَلَا تَحْنَثْ إِنَّا وَجَدْنَٰهُ صَابِرًا نِّعْمَ ٱلْعَبْدُ إِنَّهُۥٓ أَوَّابٌ.

السؤال: من هو النبي المعروف بالصبر؟

Question: Who was the prophet that was known for patience?

Predicted Answers	Gold Answer
• أَيُّوبَ • واذكر عبدنا أيوب • عبدنا أيوب • ...	• أَيُّوبَ

(c)

Figure 5.5. Three success examples of single-answer questions correctly answered by CL-AraBERT$_{\text{msa}+qrcd}$. Text highlighted in blue represent reference expressions to the respective preceding antecedents highlighted in yellow.

the next points below.

(2) **Multi-verse reasoning**. Many questions require multi-verse/sentence reasoning and coreference resolution to extract the correct answer span. In Figure 5.4(a), we speculate that our reader failed to correctly answer the *why* question because it requires multi-verse reasoning. Also, the presence of the common word ("al-jub" in Arabic, which means "a well" in English) between the question and the wrongly predicted answer could have provided a false clue. On the other hand, the reader seems to have succeeded in applying multi-verse reasoning and coreference resolution to answer the two factoid questions in Figures 5.5(a) and 5.5(c), despite the relatively large distance between the antecedents (high-

lighted in yellow) and the reference expressions (highlighted in blue) in the respective Qur'anic passages; the distance reached 2 verses with $\sim$78 words for the anaphoric (i.e., coreference) expression in Figure 5.5(a), and 2 verses with $\sim$33 words for the expression in Figure 5.5(c).

(3) **Vocabulary mismatch**. The classical challenge of vocabulary mismatch between the question and answer vocabularies has also contributed to some failure incidences. For the multi-answer question in Figure 5.2(b), the reader failed to return the first gold answer component (probably due to the absence of any term overlap), but interestingly, it was able to return the second answer component despite the absence of any term overlap.

Another interesting example is demonstrated in Figure 5.4(b), where the reader failed to answer the single-answer question not only due to the absence of term overlap, but also due to the nature of the answer being evidence-based (as mentioned earlier); however, the reader was able to return the answer term ("al-tayr" in Arabic, which means "a bird" in English) by associating it to the question term ("al-haywanat" in Arabic, which means "animals" in English). This could be considered an implicit form of query expansion. Moreover, Figure 5.5(b), demonstrates another vivid example of implicit query expansion, where the reader has successfully returned the two occurrences of the gold answer ("al-naciya" or "naciya" in Arabic, which means "forepart of the head" in English) to the single-answer question, despite the absence of any term overlap between the question and the gold answer terms.

Finally, Figure 5.3(a) showcases the reader's ability to successfully answer the *why* question by conquering both challenges, the vocabulary mismatch challenge and the evidence-based nature of the answer challenge (as mentioned under the first challenge above).

(4) **Incorrect verse context**. Another challenge is predicting a gold-matching answer span that is not extracted from the gold verse intended, i.e., the context verse that belongs to the original verse-based *direct* answer(s) to which the annotators extracted the gold answer spans from. We recall that the *direct* answers were initially annotated, based on their contexts, by Qur'an experts while developing *AyaTEC* [85]. As such, the adopted evaluation measures will not reward a system/model for predicting such an answer given that the answer matching function is based on token positions, as explained in Section 3.2.3.1. For the factoid and single-answer question in Figure 5.4(c), the reader returned the wrong occurrence of the gold answer (highlighted in pink) that is located outside the correct gold context that includes the coreference expression (highlighted in blue) to the antecedent, which happens to be the gold answer (highlighted in yellow).

(5) **Partial failures**. There were also some partial failures due to one or more of the following reasons: i) not predicting all the answer components of a multi-answer question (e.g., missing the third gold answer component in Figure 3.9); ii) partially predicting an answer, while leaving out an essential word/phrase (e.g., the first predicted answer in Figure 5.4(b)); or iii) predicting an answer span that includes a non-essential word/phrase (e.g., the second predicted answers in Figure 3.9 and Figure 5.2(b)).

As a future direction to enhance performance over multi-answer questions, we may consider casting the reading comprehension task as a sequence tagging problem to increase the probability of predicting and discovering all the answer components.

Another future direction to enhance multi-verse reasoning, over both question types, is to improve coreference resolution by exploiting the QurAna corpus by Sharaf and Atwell [121], which is a large corpus of the Qur'an annotated with pronominal anaphora.

CHAPTER 6: END-TO-END QA SYSTEM ON THE HOLY QUR'AN AND GENERAL IMPLICATIONS

In this chapter, we describe our approach in integrating the retriever and reader components to constitute our complete end-to-end machine reading at scale QA system on the Holy Qur'an. Given a question in MSA, our QA system should return a ranked list of answers (spans) from the Holy Qur'an.

This chapter is composed of three main sections. The first section presents an overview of the retriever-reader architecture of the QA system before shedding more light on the integration procedure between the two components. The second section is dedicated to the experimental evaluation of the QA system, where we describe the experimental setup, then present the evaluation results and discuss them in the context of addressing the last two research questions in this dissertation. This is followed by a performance analysis of the QA system. In the third section, we conclude this chapter with general implications of this research work.

6.1. The Pipelined Retriever-Reader Architecture

In Figure 6.1 (repeated again for convenience), we exhibit an overview of the pipelined retriever-reader architecture of the QA system [38], [39], [99], [148]. Given a question in MSA, the retriever component searches an inverted index of Qur'anic passages, that are expanded with two MSA resources, to help in bridging the MSA-to-CA gap. The first resource is Al-Tafseer Al-Muyassar [1], which is a simple interpretation of the Holy Qur'an in MSA, while the second is a Dictionary of Qur'anic words with their meaning in MSA [84]. The top K scoring passages that are returned by the Okapi BM25 [113] index search are then passed to the fine-tuned CL-AraBERT reader

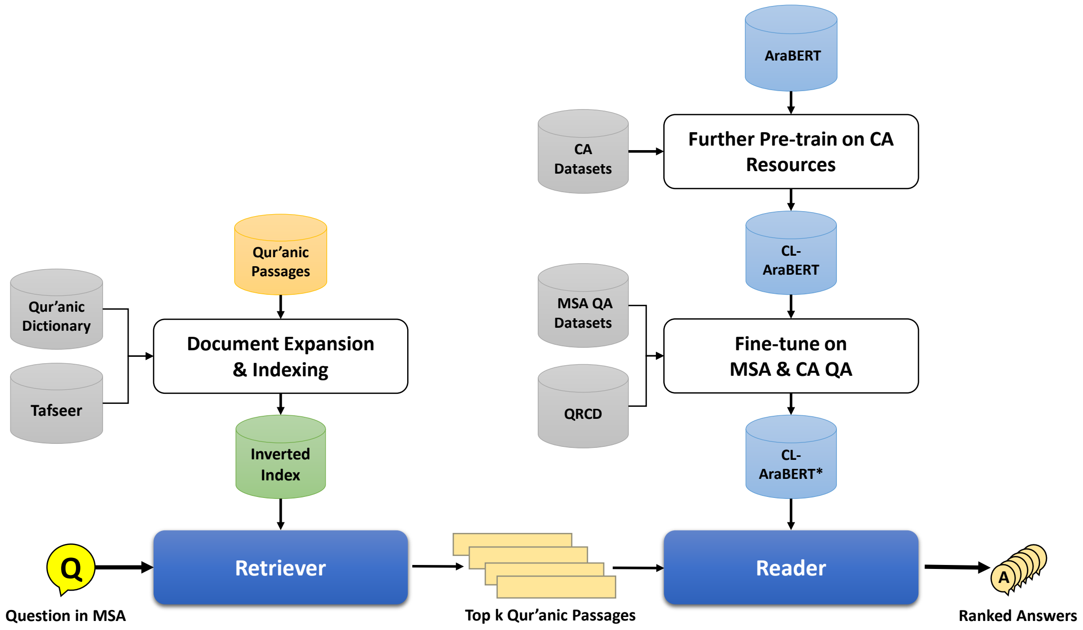


Figure 6.1. An overview of the pipelined Retriever-Reader architecture of the QA System.

as Qur'anic-only passages (i.e., after stripping the MSA text from them). The reader in turn extracts and returns the best answers from *all* these passages ranked by their normalized score.

The reader was developed by first further pre-training AraBERT [23] using about 1.05B-word Classical Arabic corpus to complement the MSA resources used in pre-training the initial model, and make it a better fit for our task. Finally, we fine-tuned CL-AraBERT as a reader using two MRC datasets in MSA, prior to fine-tuning it using our *QRCD* dataset. We cast the problem as a cross-lingual transfer learning task from MSA to CA not only to address the MSA-to-CA gap, but also to overcome the modest size of the *QRCD* dataset.

Integrating the Retriever and Reader Components

To integrate the retriever and reader components, we reformat the search hit list of the top K passages resulting from an index search by the retriever (for a set of questions), into a BERT-compliant input format to be fed to our best performing CL-AraBERT reader (CL-AraBERT$_{\text{msa}+qrcd}$). The reader in turn predicts the top R answer spans from each question-passage pair at a time. For the predicted answer scores to be comparable across passages, it was important to remove the softmax layer (as suggested in [41], [135]) to allow for aggregation and normalization, i.e., rather than applying the softmax on the start/end logits of predicted answers over all the words in the accompanying passage only, we delay the normalization of the softmax function, such that it is applied over the top R predicted answers extracted from *all* the top K retrieved passages for a given question (i.e., the normalization is applied over $R \times K$ predicted answers). Finally, the reader returns the re-ranked predicted answers by their normalized scores.

6.2. Evaluating the End-to-End QA System

In this section, we describe the setup of our experiments, then present the evaluation results and discuss them in the context of addressing the fifth and sixth research questions listed below in **black**.

RQ1: Would expanding the Qur'anic passages with their corresponding Qur'an related MSA resources help the retriever in bridging the gap between the questions in MSA and their answer-bearing Qur'anic passages?

RQ2: Does further pre-training with Classical Arabic improve the performance over the MSA-only pre-trained model?

RQ3: Would it be enough to exclusively rely on transfer learning from MSA to CA in fine-tuning the readers without the need for MRC datasets in Classical Arabic?

RQ4: Adopting the passage-scope for evaluation, how does the fine-tuned CL-AraBERT reader perform on multi-answer questions vs. single-answer questions?

RQ5: Adopting the Qur'an-scope for evaluation, how does the end-to-end QA system perform on multi-answer questions vs. single-answer questions?

RQ6: Is a native BERT-based model architecture fine-tuned as an extractive MRC reader sub-optimal for QA and MRC tasks over multi-answer questions?

6.2.1. Experimental Setup

We evaluate the QA system on the holdout dataset that was randomly split over the unique questions in *QRCD* as depicted in Table 4.1. The holdout dataset is composed of 34 unique questions; 13 of which are single-answer questions, while the remaining 21 are multi-answer questions. We note that the same holdout dataset (with the same random split/seed) was used for evaluating the retriever component and the reader component in Sections 4.6 and 5.2, respectively. Though for the reader, the distribution of questions and their question-passage-triplets are based on the passage-scope (rather than the Qur'an-scope), where each question-passage occurrence was considered an independent question as shown in Table 5.1. In essence, the holdout experimental setup for evaluating the reader in Section 5.2.1 was also adopted for evaluating the end-to-end QA system, but at the Qur'an-scope rather than the passage-scope.

Since fine-tuning the CL-AraBERT reader was performed five times with a different random seed for each run in the holdout setup (as described in Section 5.2.1), we evaluated the performance of the QA system five times as well. In each evaluation run, we coupled the retriever with one of the five fine-tuned CL-AraBERT readers. The median performance among the five runs was reported per evaluation metric over all questions. As indicated in Section 3.2.3, Partial Average Precision (*pAP*) was the rank-based measure used for evaluating multi-answer and single-answer questions, whereas $F_1@1$ and *EM* were the set-based measures used for evaluating single-answer questions only. We have also used Partial Reciprocal Rank pRR as a rank-based measure for evaluating single-answer questions (as described in Section 3.1.6.2) for an experiment to address RQ6.

6.2.2. Results and Discussion

Table 6.1. Results of the end-to-end QA system across question types in the QRCD dataset. The top R answers from the top K passages are considered in the evaluation. The letters "S" and "M" correspond to "single-answer" and "multi-answer" questions, respectively.

Top K	Top R	Question	QRCD Test / Holdout (Qur'an-scope)		
Passages	Answers	Type	$F_1@1$	EM	$pAP@10$
20	1	S	21.42	7.69	27.61
		M			13.63
		All			19.35
20	2	S	21.42	7.69	27.90
		M			13.42
		All			18.92
20	3	S	21.42	7.69	27.88
		M			13.34
		All			18.77

With about 80% of the unique questions in *QRCD* comprising two or more answers, it was essential to address RQ5 that is concerned with comparing the performance of the QA system across question types. Table 6.1 presents the comparison in terms of all possible evaluation measures. We evaluate the system over the answers predicted from the first, second and third best answers extracted from the top 20 retrieved Qur'anic passages. To answer RQ5, the results clearly show that the system performed better, in terms of pAP, on single-answer questions in comparison to multi-answer questions. This is expected since multi-answer questions are naturally more challenging, hence typically harder. The attained pAP scores on single-answer questions were better than those on multi-answer questions by 13.98 points when the top first answer from each of the retrieved passages were considered in the evaluation. Considering more answers from the retrieved passages (second and third part in Table 6.1) did not seem to help in enhancing the pAP score on multi-answer questions; in fact, it witnessed a marginal deterioration.

Table 6.2. Results of evaluating multi-answer questions as single-answer questions by the end-to-end QA system. Only the top answers from the top K passages are considered in the evaluation. The letters "S" and "M" correspond to "single-answer" and "multi-answer" questions, respectively.

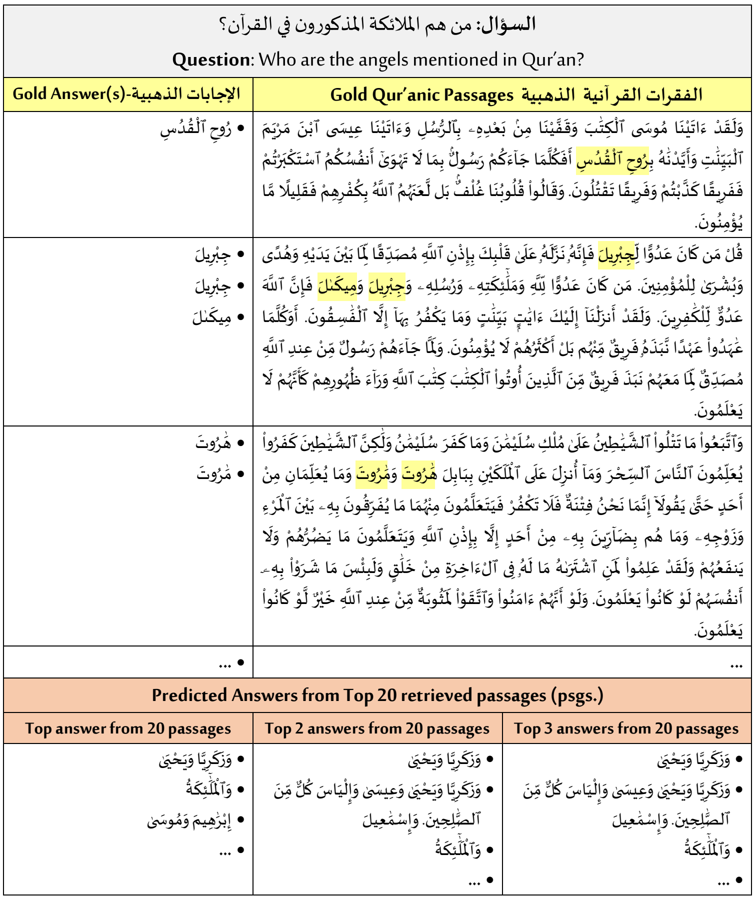
Top K	Top R	Question	QRCD Test / Holdout (Qur'an-scope)		
Passages	Answers	Type	F_1@1	EM	pRR
20	1	S	21.42	7.69	27.61
		M	22.18	9.52	26.55
		All	23.60	8.82	26.94

The above finding may suggest (along with insight drawn from the performance analysis of the reader in isolation of the retriever component 5.2.2.4) that a native BERT-based model architecture fine-tuned on the MRC task may not be intrinsically optimal for multi-answer questions. To gather more evidence on this finding, we evaluated multi-answer questions as single-answer questions and rewarded the system for retrieving *any* answer component. Instead of using pAP, we used Partial Reciprocal Rank pRR as an alternative rank-based measure that is more suitable for evaluating single-answer questions (as described in Section 3.1.6.2). The results in Table 6.2 show that the reader's performance over pseudo single-answer questions (i.e., multi-answer questions) attained comparable scores to (if not sometimes higher than) the genuine single-answer questions. To answer RQ6, we have provided enough evidence to suggest that a native BERT-based model architecture fine-tuned as an extractive MRC reader may not be optimal for the task over multi-answer questions.

In general, the witnessed overall performance of the end-to-end QA system on all questions (including single-answer questions) is modest. Similar end-to-end QA systems in the literature adopting the retriever-reader architecture (with a BERT reader), such as [141], witnessed a severe degradation in the exact match score over the SQuAD v1.1 dataset in comparison to that reported for the BERT reader in [48]. This affirms that the task is hard, but with ample room for improvement.

6.2.3. Performance Analysis of the End-to-End QA System

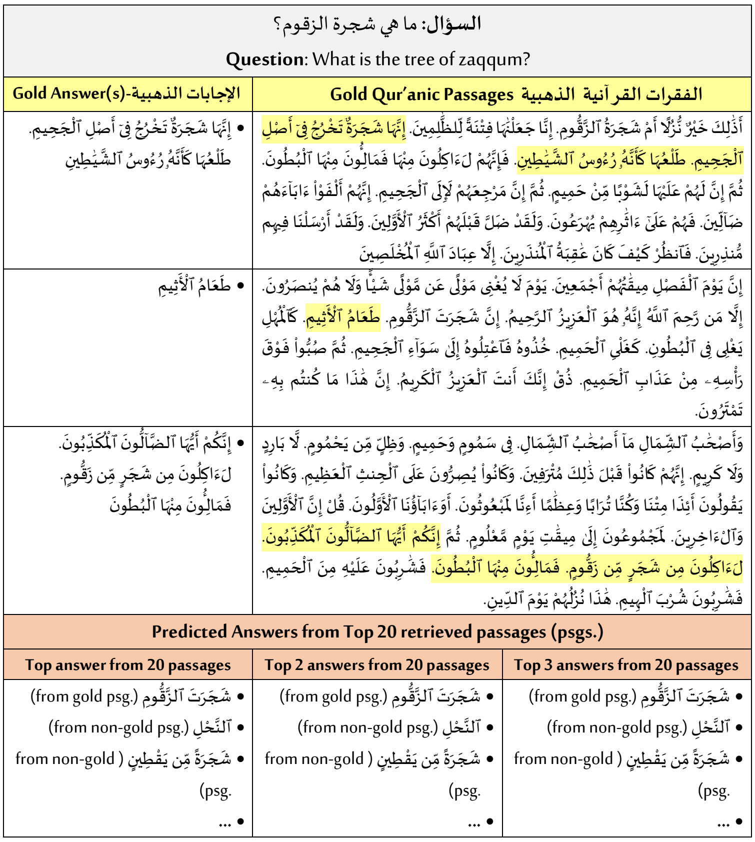
In this section, we discuss and present several failure and success examples (Figures 6.2 through 6.7) in an attempt to understand the weaknesses and strengths of the end-to-end QA system. We recall that the system is composed of the best performing retriever (expanded with Al-Tafseer and Dictionary) and the best performing reader (CL-AraBERT$_{\text{msa+}qrcd}$) on the *QRCD* dataset. This performance analysis aims at providing insights towards enhancing the modest performance of the QA system. It should not be inspected in isolation of the performance analysis of the reader (described in section 5.2.2.4) as it complements it.



<table>
<tr><th colspan="3">السؤال: من هم الملائكة المذكورون في القرآن؟
Question: Who are the angels mentioned in Qur'an?</th></tr>
<tr><th>Gold Answer(s)-الإجابات الذهبية</th><th colspan="2">Gold Qur'anic Passages الفقرات القرآنية الذهبية</th></tr>
<tr><td>• رُوحِ ٱلْقُدُسِ</td><td colspan="2">وَلَقَدْ ءَاتَيْنَا مُوسَى ٱلْكِتَٰبَ وَقَفَّيْنَا مِنۢ بَعْدِهِۦ بِٱلرُّسُلِ وَءَاتَيْنَا عِيسَى ٱبْنَ مَرْيَمَ ٱلْبَيِّنَٰتِ وَأَيَّدْنَٰهُ بِرُوحِ ٱلْقُدُسِ أَفَكُلَّمَا جَآءَكُمْ رَسُولٌۢ بِمَا لَا تَهْوَىٰٓ أَنفُسُكُمُ ٱسْتَكْبَرْتُمْ فَفَرِيقًا كَذَّبْتُمْ وَفَرِيقًا تَقْتُلُونَ. وَقَالُوا۟ قُلُوبُنَا غُلْفٌۢ بَل لَّعَنَهُمُ ٱللَّهُ بِكُفْرِهِمْ فَقَلِيلًا مَّا يُؤْمِنُونَ.</td></tr>
<tr><td>• جِبْرِيلَ
• جِبْرِيلَ
• مِيكَىٰلَ</td><td colspan="2">قُلْ مَن كَانَ عَدُوًّا لِّجِبْرِيلَ فَإِنَّهُۥ نَزَّلَهُۥ عَلَىٰ قَلْبِكَ بِإِذْنِ ٱللَّهِ مُصَدِّقًا لِّمَا بَيْنَ يَدَيْهِ وَهُدًى وَبُشْرَىٰ لِلْمُؤْمِنِينَ. مَن كَانَ عَدُوًّا لِّلَّهِ وَمَلَٰٓئِكَتِهِۦ وَرُسُلِهِۦ وَجِبْرِيلَ وَمِيكَىٰلَ فَإِنَّ ٱللَّهَ عَدُوٌّ لِّلْكَٰفِرِينَ. وَلَقَدْ أَنزَلْنَآ إِلَيْكَ ءَايَٰتٍۭ بَيِّنَٰتٍ وَمَا يَكْفُرُ بِهَآ إِلَّا ٱلْفَٰسِقُونَ. أَوَكُلَّمَا عَٰهَدُوا۟ عَهْدًا نَّبَذَهُۥ فَرِيقٌ مِّنْهُم بَلْ أَكْثَرُهُمْ لَا يُؤْمِنُونَ. وَلَمَّا جَآءَهُمْ رَسُولٌ مِّنْ عِندِ ٱللَّهِ مُصَدِّقٌ لِّمَا مَعَهُمْ نَبَذَ فَرِيقٌ مِّنَ ٱلَّذِينَ أُوتُوا۟ ٱلْكِتَٰبَ كِتَٰبَ ٱللَّهِ وَرَآءَ ظُهُورِهِمْ كَأَنَّهُمْ لَا يَعْلَمُونَ.</td></tr>
<tr><td>• هَٰرُوتَ
• مَٰرُوتَ</td><td colspan="2">وَٱتَّبَعُوا۟ مَا تَتْلُوا۟ ٱلشَّيَٰطِينُ عَلَىٰ مُلْكِ سُلَيْمَٰنَ وَمَا كَفَرَ سُلَيْمَٰنُ وَلَٰكِنَّ ٱلشَّيَٰطِينَ كَفَرُوا۟ يُعَلِّمُونَ ٱلنَّاسَ ٱلسِّحْرَ وَمَآ أُنزِلَ عَلَى ٱلْمَلَكَيْنِ بِبَابِلَ هَٰرُوتَ وَمَٰرُوتَ وَمَا يُعَلِّمَانِ مِنْ أَحَدٍ حَتَّىٰ يَقُولَآ إِنَّمَا نَحْنُ فِتْنَةٌ فَلَا تَكْفُرْ فَيَتَعَلَّمُونَ مِنْهُمَا مَا يُفَرِّقُونَ بِهِۦ بَيْنَ ٱلْمَرْءِ وَزَوْجِهِۦ وَمَا هُم بِضَآرِّينَ بِهِۦ مِنْ أَحَدٍ إِلَّا بِإِذْنِ ٱللَّهِ وَيَتَعَلَّمُونَ مَا يَضُرُّهُمْ وَلَا يَنفَعُهُمْ وَلَقَدْ عَلِمُوا۟ لَمَنِ ٱشْتَرَىٰهُ مَا لَهُۥ فِى ٱلْءَاخِرَةِ مِنْ خَلَٰقٍ وَلَبِئْسَ مَا شَرَوْا۟ بِهِۦٓ أَنفُسَهُمْ لَوْ كَانُوا۟ يَعْلَمُونَ. وَلَوْ أَنَّهُمْ ءَامَنُوا۟ وَٱتَّقَوْا۟ لَمَثُوبَةٌ مِّنْ عِندِ ٱللَّهِ خَيْرٌ لَّوْ كَانُوا۟ يَعْلَمُونَ.</td></tr>
<tr><td>• ...</td><td colspan="2">...</td></tr>
<tr><th colspan="3">Predicted Answers from Top 20 retrieved passages (psgs.)</th></tr>
<tr><th>Top answer from 20 passages</th><th>Top 2 answers from 20 passages</th><th>Top 3 answers from 20 passages</th></tr>
<tr><td>• وَزَكَرِيَّا وَيَحْيَىٰ
• وَٱلْمَلَٰٓئِكَةُ
• إِبْرَٰهِيمَ وَمُوسَىٰ
• ...</td><td>• وَزَكَرِيَّا وَيَحْيَىٰ
• وَزَكَرِيَّا وَيَحْيَىٰ وَعِيسَىٰ وَإِلْيَاسَ كُلٌّ مِّنَ ٱلصَّٰلِحِينَ. وَإِسْمَٰعِيلَ
• وَٱلْمَلَٰٓئِكَةُ
• ...</td><td>• وَزَكَرِيَّا وَيَحْيَىٰ
• وَزَكَرِيَّا وَيَحْيَىٰ وَعِيسَىٰ وَإِلْيَاسَ كُلٌّ مِّنَ ٱلصَّٰلِحِينَ. وَإِسْمَٰعِيلَ
• وَٱلْمَلَٰٓئِكَةُ
• ...</td></tr>
</table>

Figure 6.2. A failure example of a multi-answer question. All incorrect answers were extracted from non-relevant (non-gold) passages.

Since the Qur'an scope is used for evaluating the end-to-end QA system, *all occurrences* of the correct answers to the questions were considered in the evaluation. This may partially explain the severe drop in the pAP scores over multi-answer questions, which is a natural consequence if the system fails to retrieve *all* the relevant (gold) answer-bearing passages to the respective questions. Surprisingly, the failure examples revealed that in many cases the retriever failed to retrieve *any* gold (answer-bearing) passages to some of the questions. For example, for the multi-answer question in Figure 6.2, no gold passages were retrieved mainly due to the vocabulary mismatch between the question and the answer vocabularies. Similarly, for the single-answer question in Figure 6.4, the gold passage was not retrieved for a different reason; it was overshadowed by false positive

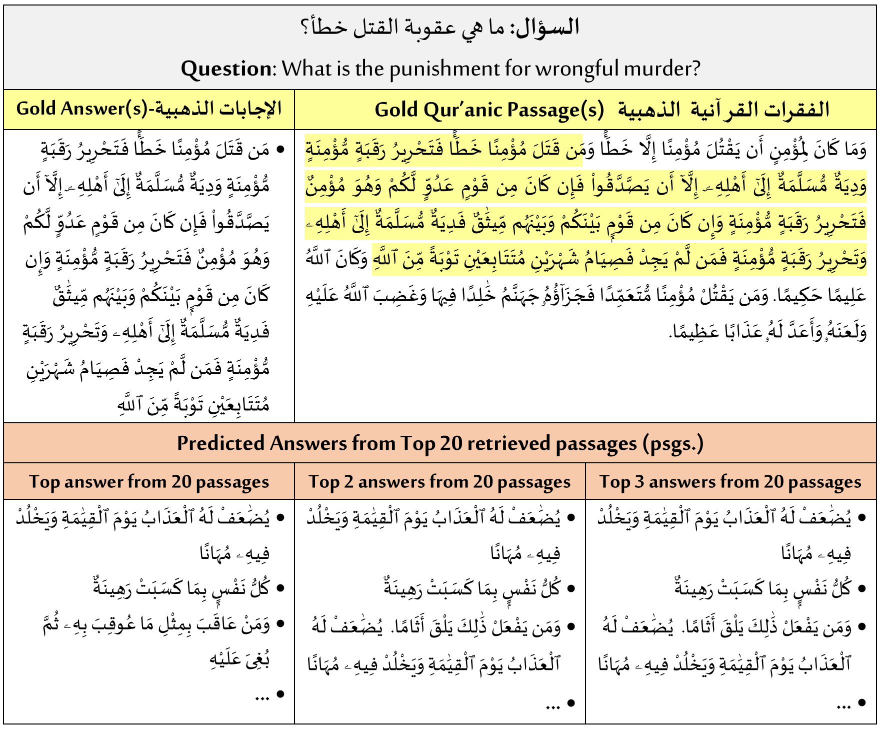


<table>
<tr><td colspan="3">السؤال: ما هي شجرة الزقوم؟
Question: What is the tree of zaqqum?</td></tr>
<tr><td>**Gold Answer(s)-الإجابات الذهبية**</td><td colspan="2">**Gold Qur'anic Passages الفقرات القرآنية الذهبية**</td></tr>
<tr><td>• إِنَّهَا شَجَرَةٌ تَخْرُجُ فِىٓ أَصْلِ ٱلْجَحِيمِ.
طَلْعُهَا كَأَنَّهُۥ رُءُوسُ ٱلشَّيَٰطِينِ</td><td colspan="2">أَذَٰلِكَ خَيْرٌ نُّزُلًا أَمْ شَجَرَةُ ٱلزَّقُّومِ. إِنَّا جَعَلْنَٰهَا فِتْنَةً لِّلظَّٰلِمِينَ. إِنَّهَا شَجَرَةٌ تَخْرُجُ فِىٓ أَصْلِ ٱلْجَحِيمِ. طَلْعُهَا كَأَنَّهُۥ رُءُوسُ ٱلشَّيَٰطِينِ. فَإِنَّهُمْ لَءَاكِلُونَ مِنْهَا فَمَالِـُٔونَ مِنْهَا ٱلْبُطُونَ. ثُمَّ إِنَّ لَهُمْ عَلَيْهَا لَشَوْبًا مِّنْ حَمِيمٍ. ثُمَّ إِنَّ مَرْجِعَهُمْ لَإِلَى ٱلْجَحِيمِ. إِنَّهُمْ أَلْفَوْا۟ ءَابَآءَهُمْ ضَآلِّينَ. فَهُمْ عَلَىٰٓ ءَاثَٰرِهِمْ يُهْرَعُونَ. وَلَقَدْ ضَلَّ قَبْلَهُمْ أَكْثَرُ ٱلْأَوَّلِينَ. وَلَقَدْ أَرْسَلْنَا فِيهِم مُّنذِرِينَ. فَٱنظُرْ كَيْفَ كَانَ عَٰقِبَةُ ٱلْمُنذَرِينَ. إِلَّا عِبَادَ ٱللَّهِ ٱلْمُخْلَصِينَ</td></tr>
<tr><td>• طَعَامُ ٱلْأَثِيمِ</td><td colspan="2">إِنَّ يَوْمَ ٱلْفَصْلِ مِيقَٰتُهُمْ أَجْمَعِينَ. يَوْمَ لَا يُغْنِى مَوْلًى عَن مَّوْلًى شَيْـًٔا وَلَا هُمْ يُنصَرُونَ. إِلَّا مَن رَّحِمَ ٱللَّهُ إِنَّهُۥ هُوَ ٱلْعَزِيزُ ٱلرَّحِيمُ. إِنَّ شَجَرَتَ ٱلزَّقُّومِ. طَعَامُ ٱلْأَثِيمِ. كَٱلْمُهْلِ يَغْلِى فِى ٱلْبُطُونِ. كَغَلْىِ ٱلْحَمِيمِ. خُذُوهُ فَٱعْتِلُوهُ إِلَىٰ سَوَآءِ ٱلْجَحِيمِ. ثُمَّ صُبُّوا۟ فَوْقَ رَأْسِهِۦ مِنْ عَذَابِ ٱلْحَمِيمِ. ذُقْ إِنَّكَ أَنتَ ٱلْعَزِيزُ ٱلْكَرِيمُ. إِنَّ هَٰذَا مَا كُنتُم بِهِۦ تَمْتَرُونَ.</td></tr>
<tr><td>• إِنَّكُمْ أَيُّهَا ٱلضَّآلُّونَ ٱلْمُكَذِّبُونَ.
لَءَاكِلُونَ مِن شَجَرٍ مِّن زَقُّومٍ.
فَمَالِـُٔونَ مِنْهَا ٱلْبُطُونَ</td><td colspan="2">وَأَصْحَٰبُ ٱلشِّمَالِ مَآ أَصْحَٰبُ ٱلشِّمَالِ. فِى سَمُومٍ وَحَمِيمٍ. وَظِلٍّ مِّن يَحْمُومٍ. لَّا بَارِدٍ وَلَا كَرِيمٍ. إِنَّهُمْ كَانُوا۟ قَبْلَ ذَٰلِكَ مُتْرَفِينَ. وَكَانُوا۟ يُصِرُّونَ عَلَى ٱلْحِنثِ ٱلْعَظِيمِ. وَكَانُوا۟ يَقُولُونَ أَئِذَا مِتْنَا وَكُنَّا تُرَابًا وَعِظَٰمًا أَءِنَّا لَمَبْعُوثُونَ. أَوَءَابَآؤُنَا ٱلْأَوَّلُونَ. قُلْ إِنَّ ٱلْأَوَّلِينَ وَٱلْءَاخِرِينَ. لَمَجْمُوعُونَ إِلَىٰ مِيقَٰتِ يَوْمٍ مَّعْلُومٍ. ثُمَّ إِنَّكُمْ أَيُّهَا ٱلضَّآلُّونَ ٱلْمُكَذِّبُونَ. لَءَاكِلُونَ مِن شَجَرٍ مِّن زَقُّومٍ. فَمَالِـُٔونَ مِنْهَا ٱلْبُطُونَ. فَشَٰرِبُونَ عَلَيْهِ مِنَ ٱلْحَمِيمِ. فَشَٰرِبُونَ شُرْبَ ٱلْهِيمِ. هَٰذَا نُزُلُهُمْ يَوْمَ ٱلدِّينِ.</td></tr>
<tr><td colspan="3">**Predicted Answers from Top 20 retrieved passages (psgs.)**</td></tr>
<tr><td>**Top answer from 20 passages**</td><td>**Top 2 answers from 20 passages**</td><td>**Top 3 answers from 20 passages**</td></tr>
<tr><td>• شَجَرَتَ ٱلزَّقُّومِ (from gold psg.)
• ٱلنَّخْلِ (from non-gold psg.)
• شَجَرَةً مِّن يَقْطِينٍ (from non-gold psg.)
• ...</td><td>• شَجَرَتَ ٱلزَّقُّومِ (from gold psg.)
• ٱلنَّخْلِ (from non-gold psg.)
• شَجَرَةً مِّن يَقْطِينٍ (from non-gold psg.)
• ...</td><td>• شَجَرَتَ ٱلزَّقُّومِ (from gold psg.)
• ٱلنَّخْلِ (from non-gold psg.)
• شَجَرَةً مِّن يَقْطِينٍ (from non-gold psg.)
• ...</td></tr>
</table>

Figure 6.3. A failure example of a multi-answer question. The first incomplete answer was partially extracted from a relevant/answer-bearing (gold) passage, while the second and third shown incorrect answers were extracted from non-relevant (non-gold) passages.

passages that have high overlap with the question but without containing the correct answer. As for the multi-answer question in Figure 6.3, the first answer was partially extracted from a gold passage, while the remaining incorrect answers were extracted from false positive passages retrieved due to some term overlap with the question.

As for the partially successful examples, Figure 6.5 exhibits a multi-answer question which does **not** have *all* its answer *components* extracted (the fourth bulleted gold answer was not among the predicted answers), nor *all* the *occurrences* of its gold answers were extracted. On the other hand, Figure 6.6 exhibits another partially successful multi-answer question that has *all* its answer *components* extracted, but **not** *all* the *occurrences* of its gold answers. Moreover, some of its returned answers



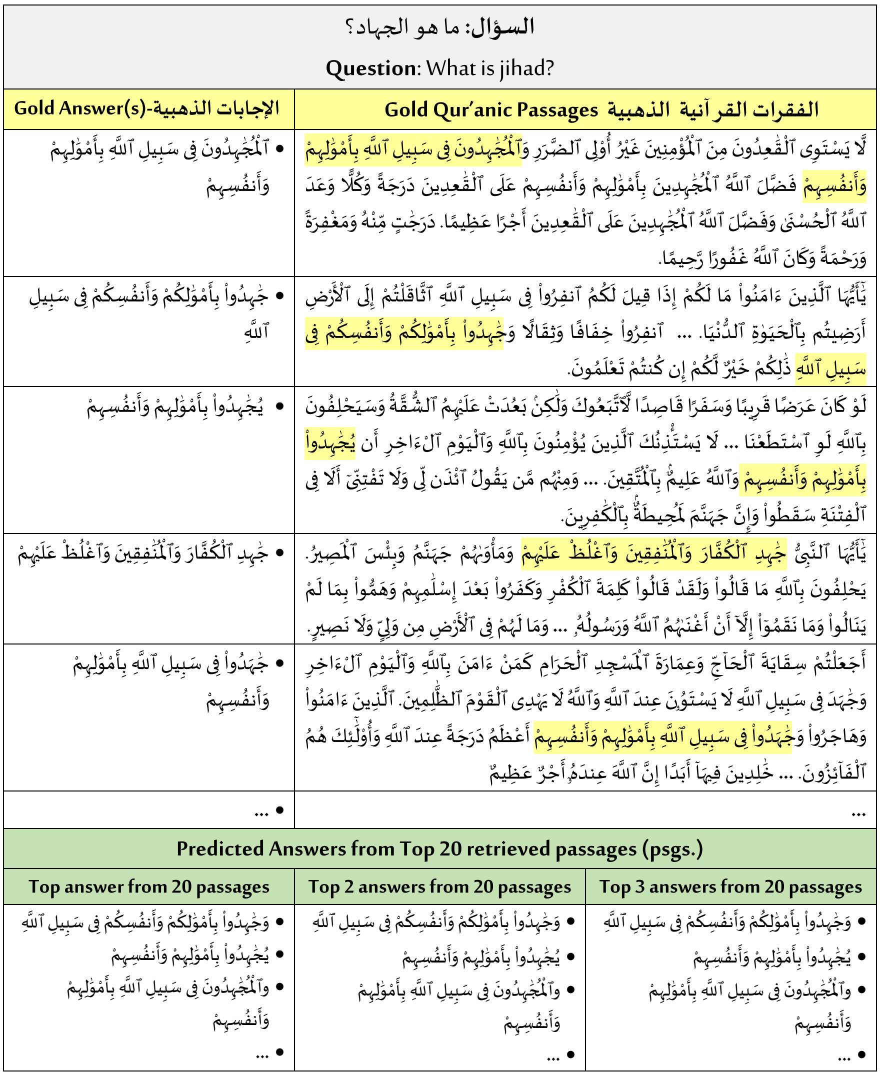
<table>
<tr><td colspan="3">السؤال: ما هي عقوبة القتل خطأ؟
Question: What is the punishment for wrongful murder?</td></tr>
<tr><td>Gold Answer(s)-الإجابات الذهبية</td><td colspan="2">Gold Qur'anic Passage(s) الفقرات القرآنية الذهبية</td></tr>
<tr><td>• مَن قَتَلَ مُؤْمِنًا خَطَـًٔا فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦٓ إِلَّآ أَن يَصَّدَّقُواْ فَإِن كَانَ مِن قَوْمٍ عَدُوٍّ لَّكُمْ وَهُوَ مُؤْمِنٌ فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَإِن كَانَ مِن قَوْمٍ بَيْنَكُمْ وَبَيْنَهُم مِّيثَٰقٌ فَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦ وَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ فَمَن لَّمْ يَجِدْ فَصِيَامُ شَهْرَيْنِ مُتَتَابِعَيْنِ تَوْبَةً مِّنَ ٱللَّهِ</td><td colspan="2">وَمَا كَانَ لِمُؤْمِنٍ أَن يَقْتُلَ مُؤْمِنًا إِلَّا خَطَـًٔا وَمَن قَتَلَ مُؤْمِنًا خَطَـًٔا فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦٓ إِلَّآ أَن يَصَّدَّقُواْ فَإِن كَانَ مِن قَوْمٍ عَدُوٍّ لَّكُمْ وَهُوَ مُؤْمِنٌ فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَإِن كَانَ مِن قَوْمٍ بَيْنَكُمْ وَبَيْنَهُم مِّيثَٰقٌ فَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦ وَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ فَمَن لَّمْ يَجِدْ فَصِيَامُ شَهْرَيْنِ مُتَتَابِعَيْنِ تَوْبَةً مِّنَ ٱللَّهِ وَكَانَ ٱللَّهُ عَلِيمًا حَكِيمًا. وَمَن يَقْتُلْ مُؤْمِنًا مُّتَعَمِّدًا فَجَزَآؤُهُۥ جَهَنَّمُ خَٰلِدًا فِيهَا وَغَضِبَ ٱللَّهُ عَلَيْهِ وَلَعَنَهُۥ وَأَعَدَّ لَهُۥ عَذَابًا عَظِيمًا.</td></tr>
<tr><td colspan="3">Predicted Answers from Top 20 retrieved passages (psgs.)</td></tr>
<tr><td>Top answer from 20 passages</td><td>Top 2 answers from 20 passages</td><td>Top 3 answers from 20 passages</td></tr>
<tr><td>• يُضَٰعَفْ لَهُ ٱلْعَذَابُ يَوْمَ ٱلْقِيَٰمَةِ وَيَخْلُدْ فِيهِۦ مُهَانًا
• كُلُّ نَفْسٍ بِمَا كَسَبَتْ رَهِينَةٌ
• وَمَنْ عَاقَبَ بِمِثْلِ مَا عُوقِبَ بِهِۦ ثُمَّ بُغِىَ عَلَيْهِ
• ...</td><td>• يُضَٰعَفْ لَهُ ٱلْعَذَابُ يَوْمَ ٱلْقِيَٰمَةِ وَيَخْلُدْ فِيهِۦ مُهَانًا
• كُلُّ نَفْسٍ بِمَا كَسَبَتْ رَهِينَةٌ
• وَمَن يَفْعَلْ ذَٰلِكَ يَلْقَ أَثَامًا. يُضَٰعَفْ لَهُ ٱلْعَذَابُ يَوْمَ ٱلْقِيَٰمَةِ وَيَخْلُدْ فِيهِۦ مُهَانًا
• ...</td><td>• يُضَٰعَفْ لَهُ ٱلْعَذَابُ يَوْمَ ٱلْقِيَٰمَةِ وَيَخْلُدْ فِيهِۦ مُهَانًا
• كُلُّ نَفْسٍ بِمَا كَسَبَتْ رَهِينَةٌ
• وَمَن يَفْعَلْ ذَٰلِكَ يَلْقَ أَثَامًا. يُضَٰعَفْ لَهُ ٱلْعَذَابُ يَوْمَ ٱلْقِيَٰمَةِ وَيَخْلُدْ فِيهِۦ مُهَانًا
• ...</td></tr>
</table>

Figure 6.4. A failure example of a single-answer question. The incorrect answers were extracted from non-relevant (non-gold) passages.

include non-essential text (e.g., the third predicted answer over matches the fourth gold answer). Finally, Figure 6.7 exhibits a single-answer question whose sole gold answer was correctly returned, but with non-essential text also included.

The above analysis has revealed the need to enhance the retriever component of the end-to-end QA system. A promising path is to adopt dense (embedding-based) passage retrieval for semantic search approaches [68], or a hybrid of both, sparse and dense retrieval approaches, as discussed in 7.2. Also, the suggestions on prospects to improve the reader (at the end of Section 5.2.2.4) are naturally among the ways to improve the end-to-end QA system.

Moreover, the analysis related to the partially successful examples, has revealed the need to tailor/adapt the measures used in the performance evaluation over questions that may have their gold answers repeated in semantically and/or syntactically similar



السؤال: ما هو الجهاد؟

Question: What is jihad?

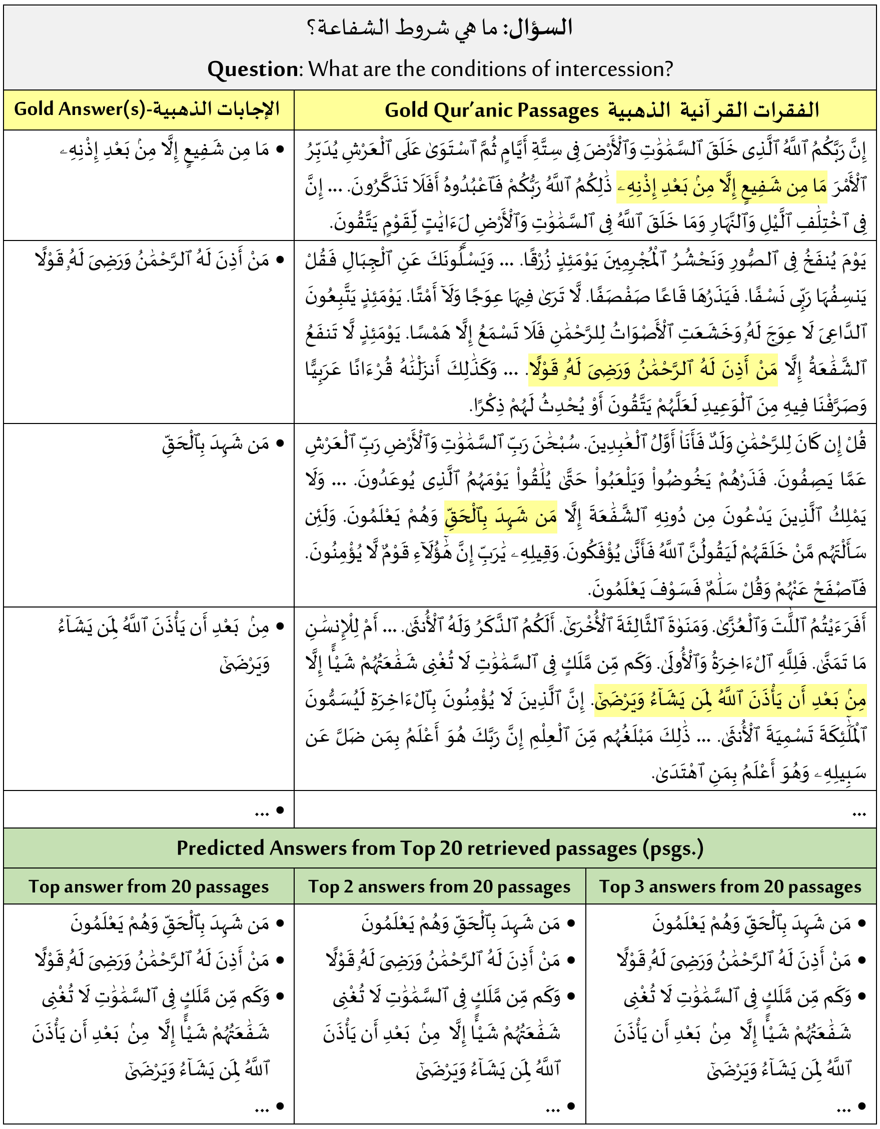
Gold Answer(s)-الإجابات الذهبية	**Gold Qur'anic Passages الفقرات القرآنية الذهبية**
• ٱلْمُجَٰهِدُونَ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ	لَّا يَسْتَوِى ٱلْقَٰعِدُونَ مِنَ ٱلْمُؤْمِنِينَ غَيْرُ أُو۟لِى ٱلضَّرَرِ وَٱلْمُجَٰهِدُونَ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ فَضَّلَ ٱللَّهُ ٱلْمُجَٰهِدِينَ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ عَلَى ٱلْقَٰعِدِينَ دَرَجَةً وَكُلًّا وَعَدَ ٱللَّهُ ٱلْحُسْنَىٰ وَفَضَّلَ ٱللَّهُ ٱلْمُجَٰهِدِينَ عَلَى ٱلْقَٰعِدِينَ أَجْرًا عَظِيمًا. دَرَجَٰتٍ مِّنْهُ وَمَغْفِرَةً وَرَحْمَةً وَكَانَ ٱللَّهُ غَفُورًا رَّحِيمًا.
• جَٰهِدُوا۟ بِأَمْوَٰلِكُمْ وَأَنفُسِكُمْ فِى سَبِيلِ ٱللَّهِ	يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوا۟ مَا لَكُمْ إِذَا قِيلَ لَكُمُ ٱنفِرُوا۟ فِى سَبِيلِ ٱللَّهِ ٱثَّاقَلْتُمْ إِلَى ٱلْأَرْضِ أَرَضِيتُم بِٱلْحَيَوٰةِ ٱلدُّنْيَا. ... ٱنفِرُوا۟ خِفَافًا وَثِقَالًا وَجَٰهِدُوا۟ بِأَمْوَٰلِكُمْ وَأَنفُسِكُمْ فِى سَبِيلِ ٱللَّهِ ذَٰلِكُمْ خَيْرٌ لَّكُمْ إِن كُنتُمْ تَعْلَمُونَ.
• يُجَٰهِدُوا۟ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ	لَوْ كَانَ عَرَضًا قَرِيبًا وَسَفَرًا قَاصِدًا لَّٱتَّبَعُوكَ وَلَٰكِنۢ بَعُدَتْ عَلَيْهِمُ ٱلشُّقَّةُ وَسَيَحْلِفُونَ بِٱللَّهِ لَوِ ٱسْتَطَعْنَا ... لَا يَسْتَـْٔذِنُكَ ٱلَّذِينَ يُؤْمِنُونَ بِٱللَّهِ وَٱلْيَوْمِ ٱلْءَاخِرِ أَن يُجَٰهِدُوا۟ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ وَٱللَّهُ عَلِيمٌۢ بِٱلْمُتَّقِينَ. ... وَمِنْهُم مَّن يَقُولُ ٱئْذَن لِّى وَلَا تَفْتِنِّىٓ أَلَا فِى ٱلْفِتْنَةِ سَقَطُوا۟ وَإِنَّ جَهَنَّمَ لَمُحِيطَةٌۢ بِٱلْكَٰفِرِينَ.
• جَٰهِدِ ٱلْكُفَّارَ وَٱلْمُنَٰفِقِينَ وَٱغْلُظْ عَلَيْهِمْ	يَٰٓأَيُّهَا ٱلنَّبِىُّ جَٰهِدِ ٱلْكُفَّارَ وَٱلْمُنَٰفِقِينَ وَٱغْلُظْ عَلَيْهِمْ وَمَأْوَىٰهُمْ جَهَنَّمُ وَبِئْسَ ٱلْمَصِيرُ. يَحْلِفُونَ بِٱللَّهِ مَا قَالُوا۟ وَلَقَدْ قَالُوا۟ كَلِمَةَ ٱلْكُفْرِ وَكَفَرُوا۟ بَعْدَ إِسْلَٰمِهِمْ وَهَمُّوا۟ بِمَا لَمْ يَنَالُوا۟ وَمَا نَقَمُوٓا۟ إِلَّآ أَنْ أَغْنَىٰهُمُ ٱللَّهُ وَرَسُولُهُۥ ... وَمَا لَهُمْ فِى ٱلْأَرْضِ مِن وَلِىٍّ وَلَا نَصِيرٍ.
• جَٰهَدُوا۟ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ	أَجَعَلْتُمْ سِقَايَةَ ٱلْحَآجِّ وَعِمَارَةَ ٱلْمَسْجِدِ ٱلْحَرَامِ كَمَنْ ءَامَنَ بِٱللَّهِ وَٱلْيَوْمِ ٱلْءَاخِرِ وَجَٰهَدَ فِى سَبِيلِ ٱللَّهِ لَا يَسْتَوُۥنَ عِندَ ٱللَّهِ وَٱللَّهُ لَا يَهْدِى ٱلْقَوْمَ ٱلظَّٰلِمِينَ. ٱلَّذِينَ ءَامَنُوا۟ وَهَاجَرُوا۟ وَجَٰهَدُوا۟ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ أَعْظَمُ دَرَجَةً عِندَ ٱللَّهِ وَأُو۟لَٰٓئِكَ هُمُ ٱلْفَآئِزُونَ. ... خَٰلِدِينَ فِيهَآ أَبَدًا إِنَّ ٱللَّهَ عِندَهُۥٓ أَجْرٌ عَظِيمٌ
• ...	...

Predicted Answers from Top 20 retrieved passages (psgs.)

Top answer from 20 passages	**Top 2 answers from 20 passages**	**Top 3 answers from 20 passages**
• وَجَٰهِدُوا۟ بِأَمْوَٰلِكُمْ وَأَنفُسِكُمْ فِى سَبِيلِ ٱللَّهِ • يُجَٰهِدُوا۟ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • وٱلْمُجَٰهِدُونَ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • ...	• وَجَٰهِدُوا۟ بِأَمْوَٰلِكُمْ وَأَنفُسِكُمْ فِى سَبِيلِ ٱللَّهِ • يُجَٰهِدُوا۟ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • وٱلْمُجَٰهِدُونَ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • ...	• وَجَٰهِدُوا۟ بِأَمْوَٰلِكُمْ وَأَنفُسِكُمْ فِى سَبِيلِ ٱللَّهِ • يُجَٰهِدُوا۟ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • وٱلْمُجَٰهِدُونَ فِى سَبِيلِ ٱللَّهِ بِأَمْوَٰلِهِمْ وَأَنفُسِهِمْ • ...

Figure 6.5. A partially successful example of a multi-answer question. Not all answer components or occurrences were extracted and returned. The three dots "**...**" in paragraphs indicate omitted text for space considerations.

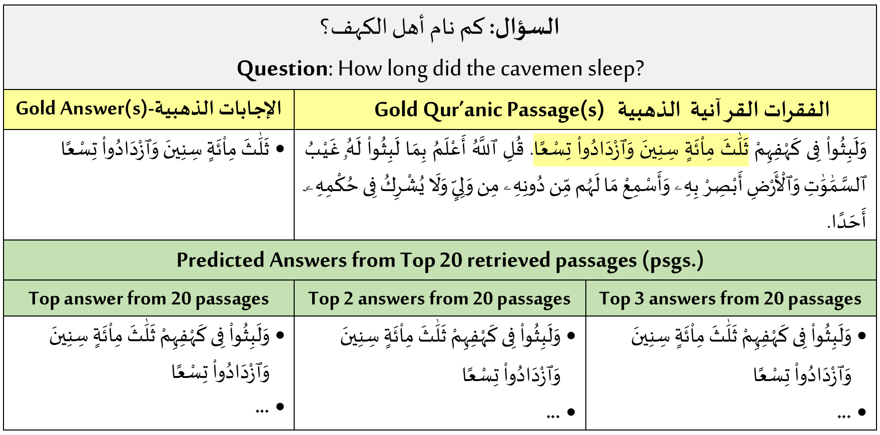
forms in more than one location in the Qur'an. In essence, the evaluation should be adapted to support two user satisfaction scenarios. In the first scenario, the user would be satisfied to get any one *occurrence* of an answer to his/her question from the system; as such, the repeated occurrences of the answer can be ignored in the evaluation. In the second scenario (which is the scenario adopted above), the user would anticipate getting all occurrences of an answer to his/her question. We note that the evaluation measures proposed for the *AyaTEC* dataset in section 3.1.6 were designed to cater for both user



<table>
<tr><th colspan="3">السؤال: ما هي شروط الشفاعة؟
Question: What are the conditions of intercession?</th></tr>
<tr><th>Gold Answer(s)-الإجابات الذهبية</th><th colspan="2">Gold Qur'anic Passages الفقرات القرآنية الذهبية</th></tr>
<tr><td>• مَا مِن شَفِيعٍ إِلَّا مِنۢ بَعْدِ إِذْنِهِۦ</td><td colspan="2">إِنَّ رَبَّكُمُ ٱللَّهُ ٱلَّذِى خَلَقَ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضَ فِى سِتَّةِ أَيَّامٍ ثُمَّ ٱسْتَوَىٰ عَلَى ٱلْعَرْشِ يُدَبِّرُ ٱلْأَمْرَ مَا مِن شَفِيعٍ إِلَّا مِنۢ بَعْدِ إِذْنِهِۦ ذَٰلِكُمُ ٱللَّهُ رَبُّكُمْ فَٱعْبُدُوهُ أَفَلَا تَذَكَّرُونَ. ... إِنَّ فِى ٱخْتِلَٰفِ ٱلَّيْلِ وَٱلنَّهَارِ وَمَا خَلَقَ ٱللَّهُ فِى ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ لَءَايَٰتٍ لِّقَوْمٍ يَتَّقُونَ.</td></tr>
<tr><td>• مَنْ أَذِنَ لَهُ ٱلرَّحْمَٰنُ وَرَضِىَ لَهُۥ قَوْلًا</td><td colspan="2">يَوْمَ يُنفَخُ فِى ٱلصُّورِ وَنَحْشُرُ ٱلْمُجْرِمِينَ يَوْمَئِذٍ زُرْقًا. ... وَيَسْـَٔلُونَكَ عَنِ ٱلْجِبَالِ فَقُلْ يَنسِفُهَا رَبِّى نَسْفًا. فَيَذَرُهَا قَاعًا صَفْصَفًا. لَّا تَرَىٰ فِيهَا عِوَجًا وَلَآ أَمْتًا. يَوْمَئِذٍ يَتَّبِعُونَ ٱلدَّاعِىَ لَا عِوَجَ لَهُۥ وَخَشَعَتِ ٱلْأَصْوَاتُ لِلرَّحْمَٰنِ فَلَا تَسْمَعُ إِلَّا هَمْسًا. يَوْمَئِذٍ لَّا تَنفَعُ ٱلشَّفَٰعَةُ إِلَّا مَنْ أَذِنَ لَهُ ٱلرَّحْمَٰنُ وَرَضِىَ لَهُۥ قَوْلًا. ... وَكَذَٰلِكَ أَنزَلْنَٰهُ قُرْءَانًا عَرَبِيًّا وَصَرَّفْنَا فِيهِ مِنَ ٱلْوَعِيدِ لَعَلَّهُمْ يَتَّقُونَ أَوْ يُحْدِثُ لَهُمْ ذِكْرًا.</td></tr>
<tr><td>• مَن شَهِدَ بِٱلْحَقِّ</td><td colspan="2">قُلْ إِن كَانَ لِلرَّحْمَٰنِ وَلَدٌ فَأَنَا۠ أَوَّلُ ٱلْعَٰبِدِينَ. سُبْحَٰنَ رَبِّ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ رَبِّ ٱلْعَرْشِ عَمَّا يَصِفُونَ. فَذَرْهُمْ يَخُوضُوا۟ وَيَلْعَبُوا۟ حَتَّىٰ يُلَٰقُوا۟ يَوْمَهُمُ ٱلَّذِى يُوعَدُونَ. ... وَلَا يَمْلِكُ ٱلَّذِينَ يَدْعُونَ مِن دُونِهِ ٱلشَّفَٰعَةَ إِلَّا مَن شَهِدَ بِٱلْحَقِّ وَهُمْ يَعْلَمُونَ. وَلَئِن سَأَلْتَهُم مَّنْ خَلَقَهُمْ لَيَقُولُنَّ ٱللَّهُ فَأَنَّىٰ يُؤْفَكُونَ. وَقِيلِهِۦ يَٰرَبِّ إِنَّ هَٰٓؤُلَآءِ قَوْمٌ لَّا يُؤْمِنُونَ. فَٱصْفَحْ عَنْهُمْ وَقُلْ سَلَٰمٌ فَسَوْفَ يَعْلَمُونَ.</td></tr>
<tr><td>• مِنۢ بَعْدِ أَن يَأْذَنَ ٱللَّهُ لِمَن يَشَآءُ وَيَرْضَىٰٓ</td><td colspan="2">أَفَرَءَيْتُمُ ٱللَّٰتَ وَٱلْعُزَّىٰ. وَمَنَوٰةَ ٱلثَّالِثَةَ ٱلْأُخْرَىٰٓ. أَلَكُمُ ٱلذَّكَرُ وَلَهُ ٱلْأُنثَىٰ. ... أَمْ لِلْإِنسَٰنِ مَا تَمَنَّىٰ. فَلِلَّهِ ٱلْءَاخِرَةُ وَٱلْأُولَىٰ. وَكَم مِّن مَّلَكٍ فِى ٱلسَّمَٰوَٰتِ لَا تُغْنِى شَفَٰعَتُهُمْ شَيْـًٔا إِلَّا مِنۢ بَعْدِ أَن يَأْذَنَ ٱللَّهُ لِمَن يَشَآءُ وَيَرْضَىٰٓ. إِنَّ ٱلَّذِينَ لَا يُؤْمِنُونَ بِٱلْءَاخِرَةِ لَيُسَمُّونَ ٱلْمَلَٰٓئِكَةَ تَسْمِيَةَ ٱلْأُنثَىٰ. ... ذَٰلِكَ مَبْلَغُهُم مِّنَ ٱلْعِلْمِ إِنَّ رَبَّكَ هُوَ أَعْلَمُ بِمَن ضَلَّ عَن سَبِيلِهِۦ وَهُوَ أَعْلَمُ بِمَنِ ٱهْتَدَىٰ.</td></tr>
<tr><td>• ...</td><td colspan="2">...</td></tr>
<tr><th colspan="3">Predicted Answers from Top 20 retrieved passages (psgs.)</th></tr>
<tr><th>Top answer from 20 passages</th><th>Top 2 answers from 20 passages</th><th>Top 3 answers from 20 passages</th></tr>
<tr><td>• مَن شَهِدَ بِٱلْحَقِّ وَهُمْ يَعْلَمُونَ
• مَنْ أَذِنَ لَهُ ٱلرَّحْمَٰنُ وَرَضِىَ لَهُۥ قَوْلًا
• وَكَم مِّن مَّلَكٍ فِى ٱلسَّمَٰوَٰتِ لَا تُغْنِى شَفَٰعَتُهُمْ شَيْـًٔا إِلَّا مِنۢ بَعْدِ أَن يَأْذَنَ ٱللَّهُ لِمَن يَشَآءُ وَيَرْضَىٰٓ
• ...</td><td>• مَن شَهِدَ بِٱلْحَقِّ وَهُمْ يَعْلَمُونَ
• مَنْ أَذِنَ لَهُ ٱلرَّحْمَٰنُ وَرَضِىَ لَهُۥ قَوْلًا
• وَكَم مِّن مَّلَكٍ فِى ٱلسَّمَٰوَٰتِ لَا تُغْنِى شَفَٰعَتُهُمْ شَيْـًٔا إِلَّا مِنۢ بَعْدِ أَن يَأْذَنَ ٱللَّهُ لِمَن يَشَآءُ وَيَرْضَىٰٓ
• ...</td><td>• مَن شَهِدَ بِٱلْحَقِّ وَهُمْ يَعْلَمُونَ
• مَنْ أَذِنَ لَهُ ٱلرَّحْمَٰنُ وَرَضِىَ لَهُۥ قَوْلًا
• وَكَم مِّن مَّلَكٍ فِى ٱلسَّمَٰوَٰتِ لَا تُغْنِى شَفَٰعَتُهُمْ شَيْـًٔا إِلَّا مِنۢ بَعْدِ أَن يَأْذَنَ ٱللَّهُ لِمَن يَشَآءُ وَيَرْضَىٰٓ
• ...</td></tr>
</table>

Figure 6.6. A partially successful example of a multi-answer question. Some of the extracted answers contain non-essential text, and not all answer occurrences were extracted and returned. The three dots "**...**" in paragraphs indicate omitted text for space considerations.

satisfaction scenarios, by exploiting the additional data components developed for each multi-answer question in *AyaTEC*: an answer-instance set and a verse-to-instances map (as described in section 3.1.3.3). Similar data components can be developed for the questions in *QRCD* to facilitate the evaluation that supports the first user satisfaction scenario.



<table>
<tr><th colspan="3">السؤال: كم نام أهل الكهف؟
Question: How long did the cavemen sleep?</th></tr>
<tr><td>Gold Answer(s)-الإجابات الذهبية</td><td colspan="2">Gold Qur'anic Passage(s) الفقرات القرآنية الذهبية</td></tr>
<tr><td>• ثَلَٰثَ مِاْئَةٍ سِنِينَ وَٱزْدَادُواْ تِسْعًا</td><td colspan="2">وَلَبِثُواْ فِى كَهْفِهِمْ ثَلَٰثَ مِاْئَةٍ سِنِينَ وَٱزْدَادُواْ تِسْعًا. قُلِ ٱللَّهُ أَعْلَمُ بِمَا لَبِثُواْ لَهُۥ غَيْبُ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ أَبْصِرْ بِهِۦ وَأَسْمِعْ مَا لَهُم مِّن دُونِهِۦ مِن وَلِيٍّ وَلَا يُشْرِكُ فِى حُكْمِهِۦٓ أَحَدًا.</td></tr>
<tr><td colspan="3">Predicted Answers from Top 20 retrieved passages (psgs.)</td></tr>
<tr><td>Top answer from 20 passages</td><td>Top 2 answers from 20 passages</td><td>Top 3 answers from 20 passages</td></tr>
<tr><td>• وَلَبِثُواْ فِى كَهْفِهِمْ ثَلَٰثَ مِاْئَةٍ سِنِينَ وَٱزْدَادُواْ تِسْعًا
• ...</td><td>• وَلَبِثُواْ فِى كَهْفِهِمْ ثَلَٰثَ مِاْئَةٍ سِنِينَ وَٱزْدَادُواْ تِسْعًا
• ...</td><td>• وَلَبِثُواْ فِى كَهْفِهِمْ ثَلَٰثَ مِاْئَةٍ سِنِينَ وَٱزْدَادُواْ تِسْعًا
• ...</td></tr>
</table>

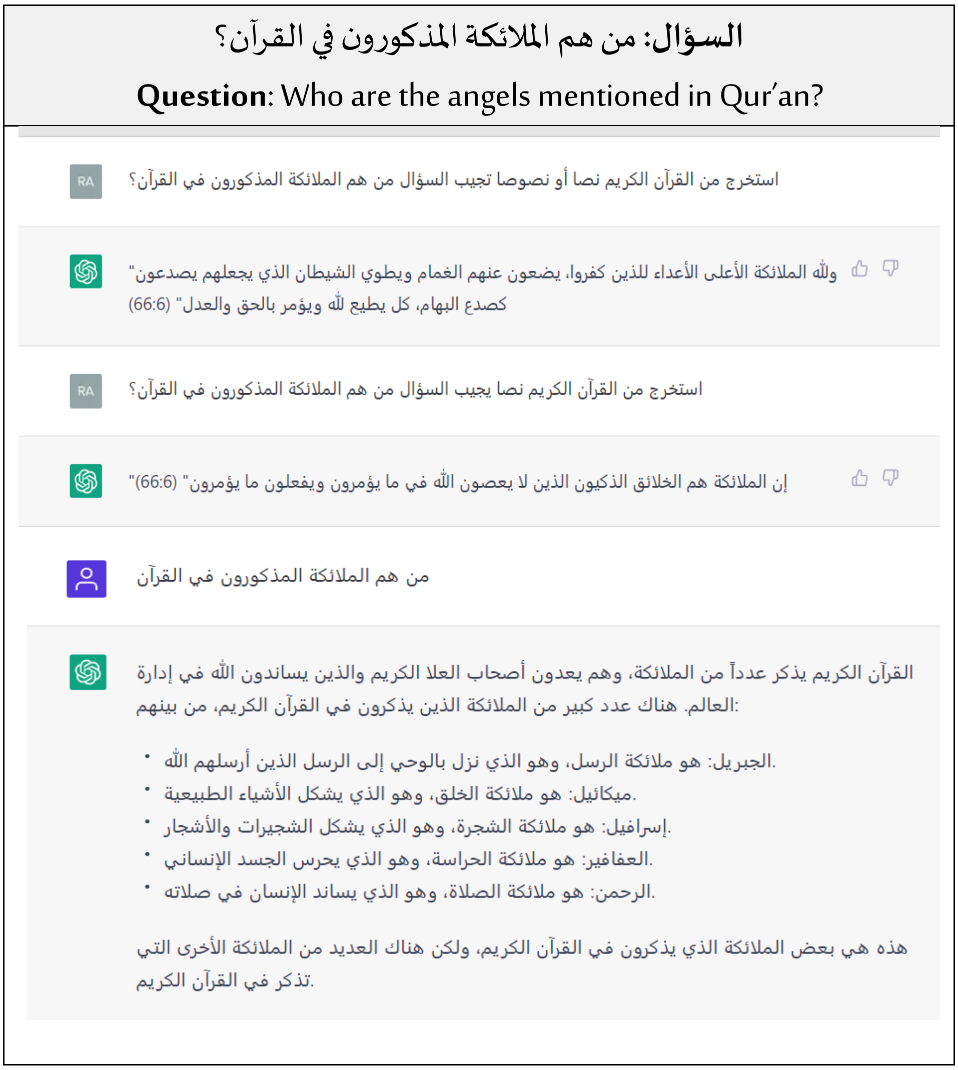
Figure 6.7. A partially successful example of a single-answer question correctly answered but with non-essential text (three words) included.

6.2.4. Exploring the Question Answering Capability of ChatGPT over the Holy Qur'an

Despite its recency, ChatGPT has gained popularity due to its remarkable (but occasionally flawed) effectiveness on a fleet of downstream tasks including question answering, extraction, summarization among others. It is a generative (decoder-only) pre-trained language model further fine-tuned using reinforcement learning from human feedback. It leveraged the same methodology adopted in training its predecessor InstructGPT [103], but utilizing a different data collection setup[1]. We perceive it as a multi-task learning model.

Out of curiosity, we showcase in Figure 6.8 through Figure 6.13, a number of QA examples using ChatGPT on the Holy Qur'an. We selected the same three questions exhibited in Figures 6.2, 6.4 and 6.7, respectively, to get a feel of the extractive QA capability of ChatGPT in comparison to our QA system. For every question, two experiments were conducted. In the first experiment, the prompts are composed of the question with an instruction to direct ChatGPT to find/extract the answer(s) from the Holy Qur'an, while in the second experiment, the prompt is composed of the question in

[1] https://openai.com/blog/chatgpt/



السؤال: من هم الملائكة المذكورون في القرآن؟

Question: Who are the angels mentioned in Qur’an?

استخرج من القرآن الكريم نصا أو نصوصا تجيب السؤال من هم الملائكة المذكورون في القرآن؟

"ولله الملائكة الأعلى الأعداء للذين كفروا، يضعون عنهم الغمام ويطوي الشيطان الذي يجعلهم يصدعون كصدع البهام، كل يطيع لله ويؤمر بالحق والعدل" (66:6)

استخرج من القرآن الكريم نصا يجيب السؤال من هم الملائكة المذكورون في القرآن؟

"إن الملائكة هم الخلائق الذكيون الذين لا يعصون الله في ما يؤمرون ويفعلون ما يؤمرون" (66:6)

من هم الملائكة المذكورون في القرآن

القرآن الكريم يذكر عدداً من الملائكة، وهم يعدون أصحاب العلا الكريم والذين يساندون الله في إدارة العالم. هناك عدد كبير من الملائكة الذين يذكرون في القرآن الكريم، من بينهم:

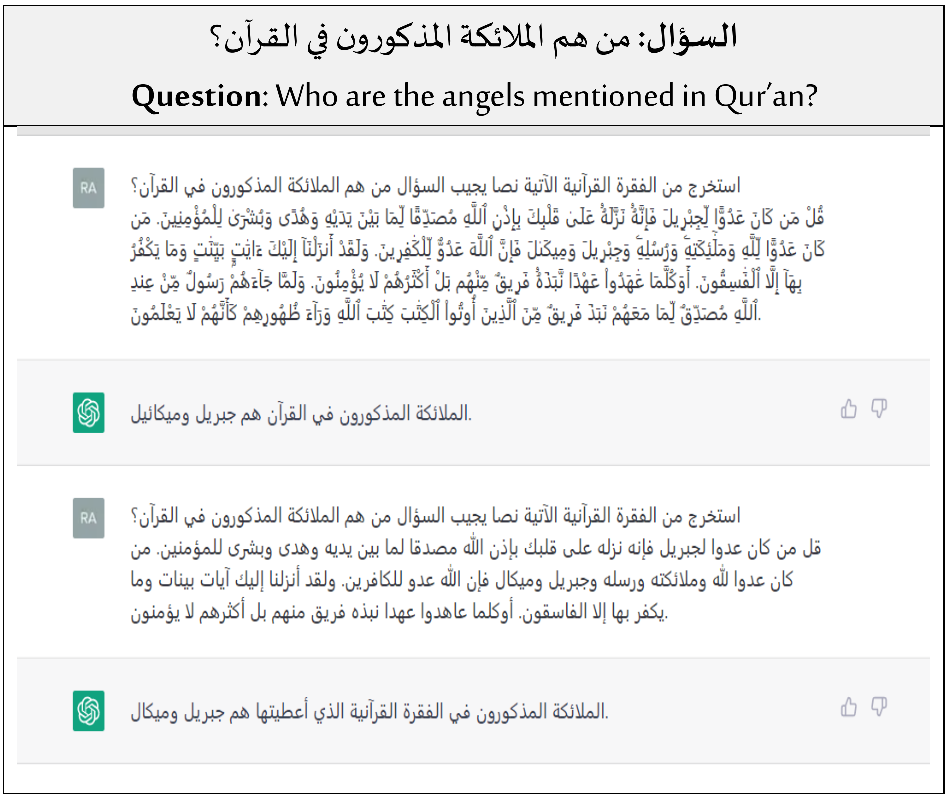
- الجبريل: هو ملائكة الرسل، وهو الذي نزل بالوحي إلى الرسل الذين أرسلهم الله.
- ميكائيل: هو ملائكة الخلق، وهو الذي يشكل الأشياء الطبيعية.
- إسرافيل: هو ملائكة الشجرة، وهو الذي يشكل الشجيرات والأشجار.
- العفافير: هو ملائكة الحراسة، وهو الذي يحرس الجسد الإنساني.
- الرحمن: هو ملائكة الصلاة، وهو الذي يساند الإنسان في صلاته.

هذه هي بعض الملائكة الذي يذكرون في القرآن الكريم، ولكن هناك العديد من الملائكة الأخرى التي تذكر في القرآن الكريم.

Figure 6.8. A failure example of a multi-answer question (from Figure 6.2) using ChatGPT over the Holy Qur’an. ChatGPT predicted flawed answers to the slightly different prompts used with the same question.

addition to the Qur’anic passage to which ChatGPT should extract the answer(s) from (thus, simulating the task of an extractive MRC Reader). Surprisingly, the behavior of ChatGPT was inconsistent with sporadic success cases and many failures.

For the multi-answer question in Figure 6.8, ChatGPT failed to list the names of the angels mentioned in the Qur’an when the prompt included the explicit instruction to *extract text from the Qur’an*. On the other hand, when the this instruction was removed, ChatGPT listed a mix of correct and wrong names of angels with some made



السؤال: من هم الملائكة المذكورون في القرآن؟

Question: Who are the angels mentioned in Qur’an?

استخرج من الفقرة القرآنية الآتية نصا يجيب السؤال من هم الملائكة المذكورون في القرآن؟
قُلْ مَن كَانَ عَدُوًّا لِّجِبْرِيلَ فَإِنَّهُۥ نَزَّلَهُۥ عَلَىٰ قَلْبِكَ بِإِذْنِ ٱللَّهِ مُصَدِّقًا لِّمَا بَيْنَ يَدَيْهِ وَهُدًى وَبُشْرَىٰ لِلْمُؤْمِنِينَ. مَن
كَانَ عَدُوًّا لِّلَّهِ وَمَلَٰٓئِكَتِهِۦ وَرُسُلِهِۦ وَجِبْرِيلَ وَمِيكَىٰلَ فَإِنَّ ٱللَّهَ عَدُوٌّ لِّلْكَٰفِرِينَ. وَلَقَدْ أَنزَلْنَآ إِلَيْكَ ءَايَٰتٍۭ بَيِّنَٰتٍ وَمَا يَكْفُرُ
بِهَآ إِلَّا ٱلْفَٰسِقُونَ. أَوَكُلَّمَا عَٰهَدُوا۟ عَهْدًا نَّبَذَهُۥ فَرِيقٌ مِّنْهُم بَلْ أَكْثَرُهُمْ لَا يُؤْمِنُونَ. وَلَمَّا جَآءَهُمْ رَسُولٌ مِّنْ عِندِ
ٱللَّهِ مُصَدِّقٌ لِّمَا مَعَهُمْ نَبَذَ فَرِيقٌ مِّنَ ٱلَّذِينَ أُوتُوا۟ ٱلْكِتَٰبَ كِتَٰبَ ٱللَّهِ وَرَآءَ ظُهُورِهِمْ كَأَنَّهُمْ لَا يَعْلَمُونَ.

الملائكة المذكورون في القرآن هم جبريل وميكائيل.

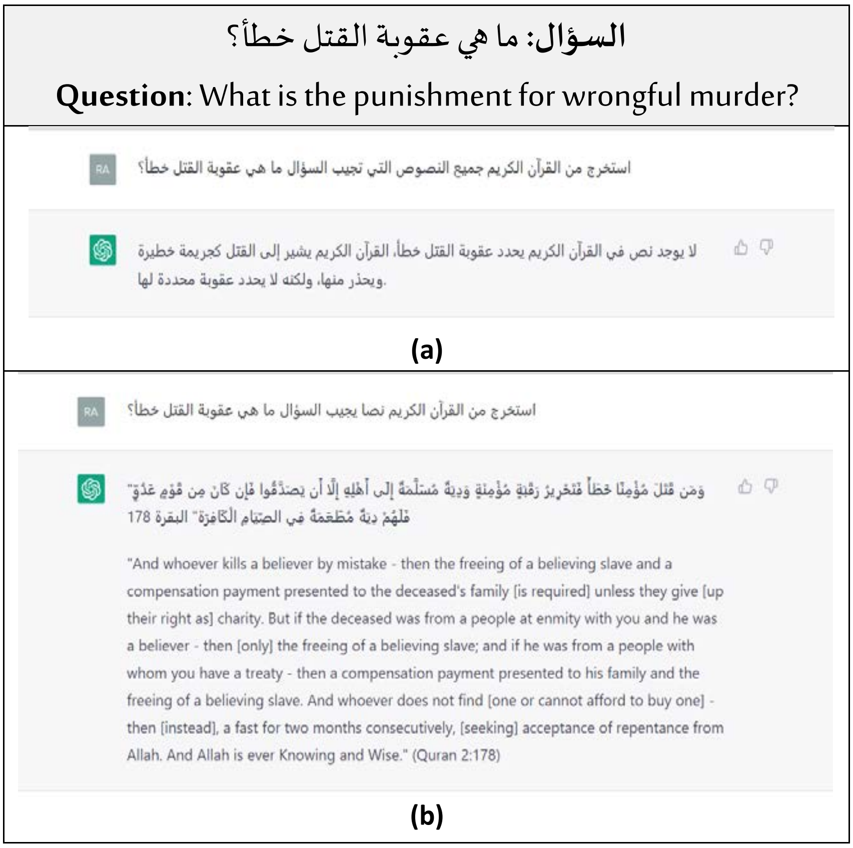
استخرج من الفقرة القرآنية الآتية نصا يجيب السؤال من هم الملائكة المذكورون في القرآن؟
قل من كان عدوا لجبريل فإنه نزله على قلبك بإذن الله مصدقا لما بين يديه وهدى وبشرى للمؤمنين. من
كان عدوا لله وملائكته ورسله وجبريل وميكال فإن الله عدو للكافرين. ولقد أنزلنا إليك آيات بينات وما
يكفر بها إلا الفاسقون. أوكلما عاهدوا عهدا نبذه فريق منهم بل أكثرهم لا يؤمنون.

الملائكة المذكورون في الفقرة القرآنية الذي أعطيتها هم جبريل وميكال.

Figure 6.9. A success example of a multi-answer question using ChatGPT over a Qur’anic passage (the second gold passage for the question in Figure 6.2). Using the Uthmani and simple-clean script versions of the Qura’nic passage did not have an effect on the predicted answer, except in the spelling of the second angel name.

up/incorrect duties. On the contrary, in Figure 6.9, ChatGPT succeeded in extracting the names of the two angels mentioned in the given Qur’anic passage. However, this success is not always warranted as ChatGPT failed to extract the name of the angel in the first gold Qur’anic passage in Figure 6.2 (example not displayed). Surprisingly, ChatGPT was be able to process the Uthmani (heavily diacritized) script as well as the simple-clean script of the Qur’anic text, seamlessly (as shown in Figure 6.9).

For the single-answer question in Figure 6.10, the prompt given to ChatGPT in the second example worked much better than the first, by predicting a partially correct answer, but with two serious problems: i) the verse label/ID is wrong; it should be Al-Nisa’a:92 instead of Al-Baqara:178, and ii) the last part of the presumably extracted



السؤال: ما هي عقوبة القتل خطأ؟

Question: What is the punishment for wrongful murder?

استخرج من القرآن الكريم جميع النصوص التي تجيب السؤال ما هي عقوبة القتل خطأ؟

لا يوجد نص في القرآن الكريم يحدد عقوبة القتل خطأ، القرآن الكريم يشير إلى القتل كجريمة خطيرة ويحذر منها، ولكنه لا يحدد عقوبة محددة لها.

(a)

استخرج من القرآن الكريم نصا يجيب السؤال ما هي عقوبة القتل خطأ؟

"وَمَن قَتَلَ مُؤْمِنًا خطأً فَتَحْرِيرُ رَقَبَةٍ مُؤْمِنَةٍ وَدِيَةٌ مُسَلَّمَةٌ إِلَى أَهْلِهِ إِلَّا أَن يَصَّدَّقُوا فَإِن كَانَ مِن قَوْمٍ عَدُوٍّ فَلَهُمْ دِيَةٌ مُطْعَمَةٌ فِي الصِّيَامِ الْكَافِرَةِ" البقرة 178

"And whoever kills a believer by mistake - then the freeing of a believing slave and a compensation payment presented to the deceased's family [is required] unless they give [up their right as] charity. But if the deceased was from a people at enmity with you and he was a believer - then [only] the freeing of a believing slave; and if he was from a people with whom you have a treaty - then a compensation payment presented to his family and the freeing of a believing slave. And whoever does not find [one or cannot afford to buy one] - then [instead], a fast for two months consecutively, [seeking] acceptance of repentance from Allah. And Allah is ever Knowing and Wise." (Quran 2:178)

(b)

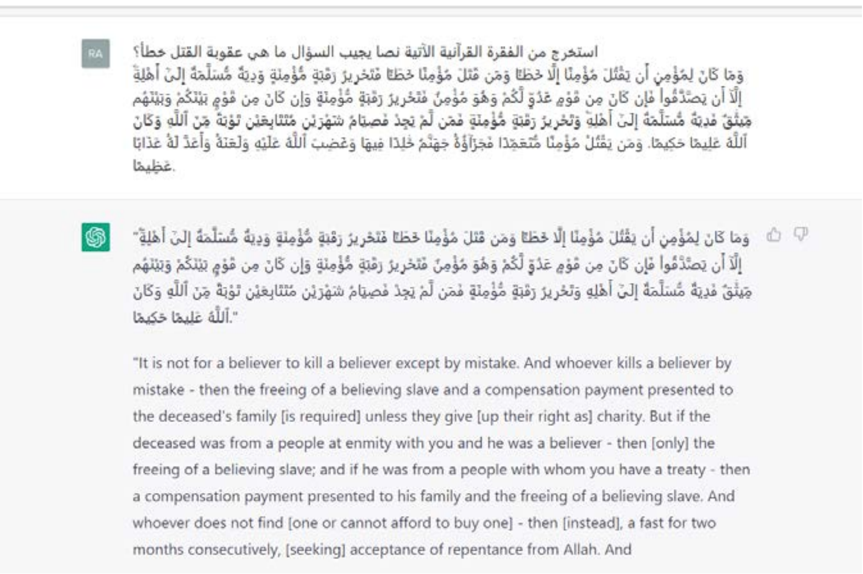
Figure 6.10. Two ChatGPT examples. (a) A failure example, and (b) a partially successful example for the same single-answer question (from Figure 6.4) using ChatGPT over the Holy Qur'an. The slight change in the instruction text included in the prompt of the second example (b) has drastically affected the predicted answer.

verse is fabricated by ChatGPT, and does not make sense! This is a serious issue, as a novice person may not know that this is not the correct verse text nor the correct verse label/ID. In Figure 6.11, ChatGPT did a better job given the Qura'nic passage. It extracted the correct answer, but with some non-essential text included. Interestingly, the English translation that appears below the answer suggest that some form of cross-lingual transfer is being deployed.

The above failure cases of ChatGPT when asked to answer questions without providing the corresponding Qur'anic passages were not surprising because they were similar to the failure cases of our end-to-end QA system. Whereas, the failure example of ChatGPT with the single-answer question in Figure 6.12 was quite surprising given

السؤال: ما هي عقوبة القتل خطأ؟

Question: What is the punishment for wrongful murder?



استخرج من الفقرة القرآنية الآتية نصا يجيب السؤال ما هي عقوبة القتل خطأ؟
وَمَا كَانَ لِمُؤْمِنٍ أَن يَقْتُلَ مُؤْمِنًا إِلَّا خَطَـًٔا وَمَن قَتَلَ مُؤْمِنًا خَطَـًٔا فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦٓ
إِلَّآ أَن يَصَّدَّقُوا۟ فَإِن كَانَ مِن قَوْمٍ عَدُوٍّ لَّكُمْ وَهُوَ مُؤْمِنٌ فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَإِن كَانَ مِن قَوْمٍۭ بَيْنَكُمْ وَبَيْنَهُم
مِّيثَـٰقٌ فَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦ وَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ فَمَن لَّمْ يَجِدْ فَصِيَامُ شَهْرَيْنِ مُتَتَابِعَيْنِ تَوْبَةً مِّنَ ٱللَّهِ وَكَانَ
ٱللَّهُ عَلِيمًا حَكِيمًا. وَمَن يَقْتُلْ مُؤْمِنًا مُّتَعَمِّدًا فَجَزَآؤُهُۥ جَهَنَّمُ خَـٰلِدًا فِيهَا وَغَضِبَ ٱللَّهُ عَلَيْهِ وَلَعَنَهُۥ وَأَعَدَّ لَهُۥ عَذَابًا
عَظِيمًا.

"وَمَا كَانَ لِمُؤْمِنٍ أَن يَقْتُلَ مُؤْمِنًا إِلَّا خَطَـًٔا وَمَن قَتَلَ مُؤْمِنًا خَطَـًٔا فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦٓ
إِلَّآ أَن يَصَّدَّقُوا۟ فَإِن كَانَ مِن قَوْمٍ عَدُوٍّ لَّكُمْ وَهُوَ مُؤْمِنٌ فَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ وَإِن كَانَ مِن قَوْمٍۭ بَيْنَكُمْ وَبَيْنَهُم
مِّيثَـٰقٌ فَدِيَةٌ مُّسَلَّمَةٌ إِلَىٰٓ أَهْلِهِۦ وَتَحْرِيرُ رَقَبَةٍ مُّؤْمِنَةٍ فَمَن لَّمْ يَجِدْ فَصِيَامُ شَهْرَيْنِ مُتَتَابِعَيْنِ تَوْبَةً مِّنَ ٱللَّهِ وَكَانَ
ٱللَّهُ عَلِيمًا حَكِيمًا."

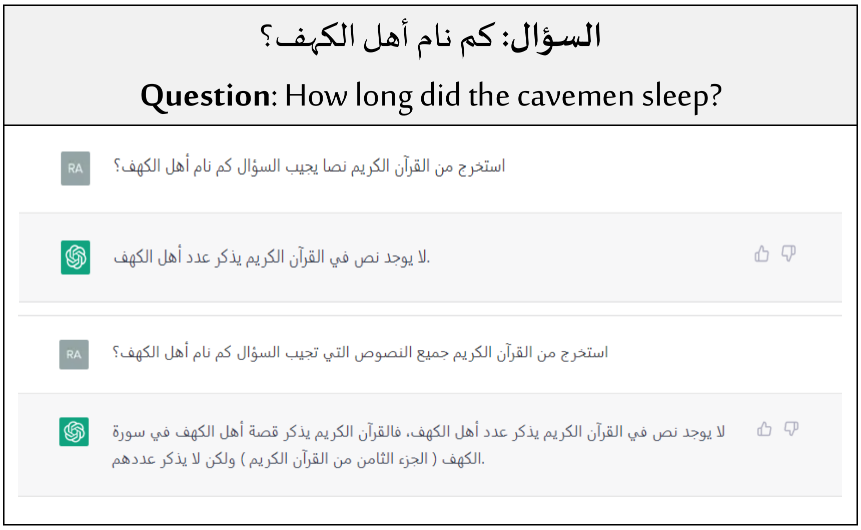
"It is not for a believer to kill a believer except by mistake. And whoever kills a believer by mistake - then the freeing of a believing slave and a compensation payment presented to the deceased's family [is required] unless they give [up their right as] charity. But if the deceased was from a people at enmity with you and he was a believer - then [only] the freeing of a believing slave; and if he was from a people with whom you have a treaty - then a compensation payment presented to his family and the freeing of a believing slave. And whoever does not find [one or cannot afford to buy one] - then [instead], a fast for two months consecutively, [seeking] acceptance of repentance from Allah. And

Figure 6.11. A partially successful example of a single-answer question using ChatGPT over a Qur'anic passage (the gold passage for the question in Figure 6.4). The answer is correct, but it includes non-essential text.

that the question is a factoid question that our QA system has correctly answered.

Moreover, With ChatGPT also succeeding (partially) in extracting the answer to the question from the given Qur'anic passage in the example of Figure 6.13, this suggests that the extractive MRC task over Qur'anic passages is relatively easier than the much harder QA task over the whole Holy Qur'an. This finding resonates well with our results and findings regarding the performance of our extractive CL-AraBERT reader and the end-to-end QA system.

The above analysis and concerns related to ChatGPT's prediction of answers with incorrect Qur'anic verses or flawed answers are an eye opener to the need for intelligent multi-entity fake detection techniques. This may include fake-verse detection, fake-fatwa detection, fake-hadith detection, among others. Such a need is of paramount importance not only due to the sensitivity of the QA task on the holy book, but also



السؤال: كم نام أهل الكهف؟

Question: How long did the cavemen sleep?

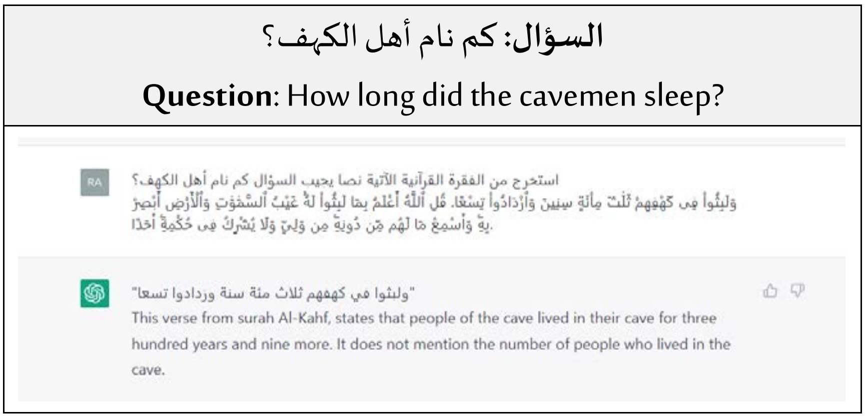
استخرج من القرآن الكريم نصا يجيب السؤال كم نام أهل الكهف؟

لا يوجد نص في القرآن الكريم يذكر عدد أهل الكهف.

استخرج من القرآن الكريم جميع النصوص التي تجيب السؤال كم نام أهل الكهف؟

لا يوجد نص في القرآن الكريم يذكر عدد أهل الكهف، فالقرآن الكريم يذكر قصة أهل الكهف في سورة
الكهف (الجزء الثامن من القرآن الكريم) ولكن لا يذكر عددهم.

Figure 6.12. A failure example of a single-answer question (from Figure 6.7) using ChatGPT over the Holy Qur'an. ChatGPT predicted flawed answers to the two slightly different prompts used with the same question.



السؤال: كم نام أهل الكهف؟

Question: How long did the cavemen sleep?

استخرج من الفقرة القرآنية الآتية نصا يجيب السؤال كم نام أهل الكهف؟
وَلَبِثُوا۟ فِى كَهْفِهِمْ ثَلَٰثَ مِا۟ئَةٍ سِنِينَ وَٱزْدَادُوا۟ تِسْعًا. قُلِ ٱللَّهُ أَعْلَمُ بِمَا لَبِثُوا۟ لَهُۥ غَيْبُ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ أَبْصِرْ
بِهِۦ وَأَسْمِعْ مَا لَهُم مِّن دُونِهِۦ مِن وَلِىٍّ وَلَا يُشْرِكُ فِى حُكْمِهِۦٓ أَحَدًا.

"ولبثوا في كهفهم ثلاث مئة سنة وردادوا تسعا"
This verse from surah Al-Kahf, states that people of the cave lived in their cave for three hundred years and nine more. It does not mention the number of people who lived in the cave.

Figure 6.13. A partially successful example of a single-answer question using ChatGPT over a Qur'anic passage (the gold passage for the question in Figure 6.7). The answer is correct, but it includes non-essential text.

as a shield against the generative rather than extractive nature of decoder-only language model architectures, such as GPT and its descendants. Also, not overlooking the risk of bias due to training these huge language models using existing resources that may include anti-Islam and anti-Qur'an content, let alone fake content.

6.3. General Implications

The research work in this dissertation has several theoretical and practical implications that we summarize below.

- **QRCD encouraging further research on the problem**. We note that the attained scores by the end-to-end QA system, the best performing CL-AraBERT reader and the retriever are relatively modest. This implies that the *QRCD* dataset is challenging enough to hopefully trigger further development of state-of-the-art QA and MRC models to enhance performance on this dataset and the task, especially for non-factoid and multi-answer questions. Moreover, being the first extractive Arabic MRC dataset on the Holy Qur'an, *QRCD* would provide a common experimental testbed for evaluating and fairly comparing the performance of future research work on this task.

- **Leveraging CL-AraBERT for other NLP CA-related tasks**. In a broader context, and based on the promising finding regarding the improvements brought upon by classical pre-training, our further pre-trained CL-AraBERT model can also be exploited for developing other NLP tasks on the Holy Qur'an and CA text, such as detecting semantic similarity between Qur'anic verses, and question answering on Hadith or Exegeses of Qur'an.

- **Facilitating partial-matching evaluation for other tasks**. On the evaluation front, we believe that the introduced *Partial Average Precision* (pAP) measure and the novel matching method (of predictions against ground truths) addresses an existing gap in the literature, not only in the context of evaluating multi-answer questions, but also in the context of evaluating other similar NLP tasks where ground truth is composed of more than one span component that might be partially-matched by the systems; e.g., the task of Named Entity Recognition (NER) in tweets. We note that the notion of partial matching, addressed in Section 3.2.3.1, can also be applied to other rank-based measures, such as $nDCG$.

- **Facilitating better understanding of the returned Qur'anic answers through knowledge enhanced QA**. Acknowledging that even native Arabic speaking Muslims may find understanding some of the Qur'anic verses quite challenging, it was important for our QA system to keep track and return with each answer the chapter and verse numbers of the Qur'anic passage(s) to which the predicted answers were extracted from. This would facilitate future enhancements on the QA system to exploit the plethora of structured and unstructured Qur'an related resources that would aid in better understanding the returned answers, such as MSA interpretations (Tafseer) of the Holy Qur'an, Hadith in addition to ontologies and knowledge bases. Exploring ways for incorporating this knowledge is a key future direction [148]. Interesting approaches to incorporate knowledge are those that exploit pre-trained language models as knowledge bases [107].

- **Prototyping the QA system as a mobile app**. To promote the practical use of the QA system and its future enhancements, it would be worthwhile to exploit mobile technology and prototype it as a mobile app. Integrating the QA system as an additional feature in mature and professional mobile apps on the Holy Qur'an could be a faster track than developing it as a Web app.

CHAPTER 7: CONCLUSION AND FUTURE WORK

We conclude this dissertation by summarizing the main findings, contributions and future directions of this research work, before listing the published and submitted publications that are related to this work.

7.1. Conclusion

In this dissertation, we have addressed the need for intelligent *machine reading at scale* over the Holy Qur'an, given the permanent interest of inquisitors and knowledge seekers in this fertile knowledge resource. We developed the pipelined retriever-reader architecture to constitute (to the best of our knowledge) the first extractive MRS QA system on the Holy Qur'an. First, a sparse passage retriever was developed over an index of Qur'anic passages expanded with Qur'an-related MSA resources to help in bridging the gap between questions in MSA and their answers in Qur'anic Classical Arabic. Second, we introduced CL-AraBERT (CLassical AraBERT), a new AraBERT-based [23] pre-trained model that is further pre-trained on about 1.05B-word Classical Arabic dataset (after being initially pre-trained on MSA datasets), to make it a better fit for NLP tasks on CA text such as the Holy Qur'an. Third, we leverage cross-lingual transfer learning from MSA to CA, and fine-tune CL-AraBERT as a reader using a couple of MSA-based MRC datasets followed by fine-tuning it on our *QRCD* dataset, to bridge the MSA-to-CA gap, and circumvent the lack of large MRC datasets in CA. Finally, the retriever-reader architecture is completed by feeding the returned top Qur'anic passages by the retriever as input to the reader for answer extraction.

We have also addressed the absence of fully-reusable QA datasets on the Holy Qur'an by first introducing *AyaTEC*, a verse-based QA dataset that we further extend to develop *QRCD*, as the first extractive Qur'anic Reading Comprehension Dataset that adopts the same format of SQuAD v1.1 [111]. Each of the two datasets serves as a common experimental test-bed to fairly compare systems, as well as a Qur'anic training resource for QA and MRC models. For the work in this dissertation, we have used *QRCD*, which is composed of 1,337 question-passage-answer triplets for 1,093 questions posed in MSA (covering both single-answer and multi-answer questions) that are coupled with their corresponding curated passages from the Qur'an. With the inclusion of multi-answer questions, *QRCD* presents an additional challenge to MRC and QA tasks.

The need to evaluate the CL-AraBERT reader and the end-to-end QA system on multi-answer questions was an eyeopener to the absence in the literature of rank-based evaluation measures that can fairly integrate partial matching for MRC and QA tasks on datasets with multi-answer questions. As such, we introduced a simple yet novel method to fairly (and partially) match the predicted answers against their respective gold answers, which we employed in the proposed *Partial Average Precision* pAP rank-based measure; pAP is an adapted version of the traditional Average Precision measure to integrate partial matching.

We have demonstrated the effective contribution of expanding the Qur'anic passages with corresponding MSA resources, in assisting the retriever to mitigate the MSA-to-CA gap.

Moreover, we empirically showed that the fine-tuned CL-AraBERT reader model significantly outperformed the similarly fine-tuned AraBERT baseline model. In general, the CL-AraBERT reader performed better on single-answer questions in comparison to multi-answer questions. Furthermore, it has also outperformed the baseline over both types of questions. However, despite the essential contribution of fine-tuning with the MSA datasets, relying exclusively on those datasets (without MRC datasets in CA, such as *QRCD*) was shown to be only sub-optimal for our reader models. This finding demonstrates the relatively high impact of the *QRCD* dataset, despite its modest size.

Performance evaluation of the CL-AraBERT reader and the end-to-end QA system were relatively modest suggesting that the MRC and QA tasks over datasets with multi-answer questions are hard. We believe there is ample room for improving their performance. As such, we make the CL-AraBERT model and the *QRCD* dataset publicly available to the research community hoping to elicit state-of-the-art research on Arabic MRC, QA and NLP on the Holy Qur'an and Classical Arabic text, such as Hadith, Exegeses of Qur'an and beyond.

We conclude with a word of caution concerning the unstructured topic diversity of the Holy Qur'an, which poses a very critical challenge to machine learning (ML) and artificial intelligence (AI) approaches, not to generate results out of their intended context. Therefore, we, as researchers, should be extra cautious of using the results of learned models without the involvement of Qur'an scholars. Bashir, Azmi, Nawaz, *et al.* [35] discuss the caveats and potential pitfalls in the Qur'anic NLP research that we should be wary of.

7.2. Future Work

Future work to enhance the performance of the QA system include several paths that could be sought with respect to the components of the retriever-reader architecture and their integration.

- **Enhancing the performance of the retriever component**. Not overlooking the significant impact of document expansion with MSA Qur'an related approaches that we have adopted, a hybrid of dense (embedding-based) and sparse bag-of-words (keyword-based) retrieval could be sought to combine the benefits of both paradigms. The simplest is to re-rank the retrieved passages returned by the BM25 index search using a neural ranker that casts the problem as a relevance classification problem [82], [101]. Alternatively, other variant architectures of dense retrieval can be used, such as representation-based [53], [68], interactive-based [100] and representation-interactive [72] retrievers. More advanced approaches such as multi-step retrieval [25], [93] are other paths to consider for addressing the challenge of multi-answer questions that may require multi-hop reasoning.

- **Enhancing the performance of the reader component**. To enhance performance over multi-answer questions, we may consider casting the reading comprehension task as a sequence tagging problem to increase the probability of predicting and discovering all the answer components. To enhance multi-verse reasoning, over both question types, coreference resolution can be improved by exploiting the QurAna corpus by Sharaf and Atwell [121], which is a large corpus of the Qur'an annotated with pronominal anaphora. Other alternatives to consider include multi-hop reasoning [142] and reinforcement learning.

  To further enhance transfer learning through pre-trained language models, we can use the more recent released versions of AraBERT (AraBERTv0.2 base and large).[1] Alternatively, other Arabic BERT-like or transformer-based models that were trained on MSA resources, such as ARBERT [7], AraELECTRA [24],

[1] https://github.com/aub-mind/arabert

AraBART [50] are worth further pre-training using the Classical Arabic corpus to compare their performance on the *QRCD* dataset with CL-AraBERT.

Moreover, using variant transformer-based models with encoder-only, decoder-only, or encoder-decoder architectures that outperformed BERT on many NLP tasks, can be another possible future direction. Some of the most prominent post-BERT models that performed well on the reading comprehension task include XLNet [141], RoBERTa [83], GPT-3 [37], BART [78], SpanBERT [66], DeBERTa [63], InstructGPT [103] and ChatGPT among others.

- **Evolving the QA system on the Holy Qur'an into a web/mobile application.** To promote the practical use of the QA system and its future enhancements, it would be worthwhile to develop it as an open source software product with an API to facilitate its future growth and presence as a web and/or mobile application.

- **Organizing another shared task for "Qur'an QA".** Planning to organize a more challenging shared task on Qur'an QA that entails developing an end-to-end QA system rather than just an MRC reader.

- **Developing a QA system on Hadith.** With the Hadith being the second source of knowledge and guidance for Muslims that complements and explains the Qur'an, it would be a natural future direction to develop a QA system on this rich resource, and an opportunity to exploit the further pre-trained CL-AraBERT. Moreover, the Hadith knowledge itself (being simpler than the Qur'an) can be used to explain the answers drawn from the QA system on Qur'an to make it a knowledge enhanced QA system (as mentioned in Section 6.3).

7.3. Published Publications

In this section, we list the related publications to this work. The first three publications are directly related to two of the core chapters in this dissertation. The first is a published journal article related to Chapter 3. The second is a published journal article, which is related to Chapter 3 and Chapter 5. The third is an overview of the shared task that we have organized in the context of OSACT 2022 workshop, which has appeared in the proceedings of the conference hosting this workshop. The overview paper is also related to Chapter 3 and Chapter 5 in addition to Section 2.5. As for the last two publications, they are on question answering research conducted during my early years of study in the PhD program.

1. **R. Malhas** and T. Elsayed, "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*, vol. 19, no. 6, pp. 1–21, Nov. 2020

2. **R. Malhas** and T. Elsayed, "Arabic Machine Reading Comprehension on the Holy Qur'an using CL-AraBERT," *Information Processing & Management*, vol. 59, no. 6, Nov. 2022

3. **R. Malhas**, W. Mansour, and T. Elsayed, "Qur'an QA 2022: Overview of the first shared task on question answering over the Holy Qur'an," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022, pp. 79–87

4. **R. Malhas**, M. Torki, and T. Elsayed, “QU-IR at SemEval 2016 Task 3: Learning to rank on Arabic community question answering forums with word embedding,” Association for Computational Linguistics (ACL), 2016

5. **R. Malhas**, M. Torki, R. Ali, T. Elsayed, and E. Yulianti, “Real, live, and concise: Answering open-domain questions with word embedding and summarization.,” in *TREC*, 2016

REFERENCES

[1] S. Aal Ash-Shaykh and Q. Scholars, *Al-Tafseer Al-Muyassar*. Medina: King Fahd Complex for the Printing of the Holy Quran, 2009.

[2] N. H. Abbas, "Quran'search for a concept'tool and website," *Unpublished thesis, University of Leeds*, 2009.

[3] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, Jun. 2016, pp. 11–16.

[4] H. Abdelbaki, M. Shaheen, and O. Badawy, "Arqa high performance Arabic question answering system," in *Arabic Language Technology International Conference (ALTIC)*, 2011, pp. 129–136.

[5] H. Abdelnasser, M. Ragab, R. Mohamed, *et al.*, "Al-Bayan: An Arabic question answering system for the holy quran," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 57–64.

[6] ——, "Al-Bayan: An Arabic question answering system for the holy quran," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 57–64.

[7] M. Abdul-Mageed, A. Elmadany, *et al.*, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7088–7105.

[8] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Idraaq: New Arabic question answering system based on query expansion and passage retrieval," in *In Proceedings of CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*, CELCT, 2012.

[9] E. Aftab and M. K. Malik, "eRock at Qur'an QA 2022: Contemporary Deep Neural Networks for Qur'an based Reading Comprehension Question Answers," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[10] B. H. Ahmed, M. K. Saad, and E. A. Refaee, "QQATeam at Quran QA 2022: Fine-Tunning Arabic QA Models for Quran QA Task," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[11] M. Akour, S. Abufardeh, K. Magel, and Q. Al-Radaideh, "Qarabpro: A rule based question answering system for reading comprehension tests in Arabic," *American Journal of Applied Sciences*, vol. 8, no. 6, pp. 652–661, 2011, ISSN: 1546-9239.

[12] A. Al Gharaibeh, A. Al Taani, and I. Alsmadi, "The usage of formal methods in quran search system," in *Proceedings of international conference on information and communication systems, Ibrid, Jordan*, Citeseer, 2011, pp. 22–24.

[13] M. Alhawarat, "Extracting topics from the holy quran using generative models," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, pp. 288–294, 2015.

[14] M. Alqahtani and E. Atwell, “Arabic quranic search tool based on ontology,” in *International Conference on Applications of Natural Language to Information Systems*, Springer, 2016, pp. 478–485.

[15] ——, “Evaluation criteria for computational quran search,” *International Journal on Islamic Applications in Computer Science And Technology*, vol. 5, no. 1, pp. 12–22, 2017.

[16] ——, “Annotated corpus of Arabic al-quran question and answer,” 2018.

[17] M. M. A. Alqahtani, “Quranic Arabic semantic search model based on ontology of concepts,” Ph.D. dissertation, University of Leeds, 2019.

[18] M. Alrabiah, A. Al-Salman, E. S. Atwell, and N. Alhelewh, “KSUCCA: A key to exploring Arabic historical linguistics,” *International Journal of Computational Linguistics (IJCL)*, vol. 5, no. 2, pp. 27–36, 2014.

[19] A. Alsaleh, S. Althabiti, I. Alshammari, *et al.*, “LK2022 at Qur’an QA 2022: Simple Transformers Model for Finding Answers to Questions from Qur’an,” in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[20] K. Alsubhi, A. Jamal, and A. Alhothali, “Pre-trained transformer-based approach for Arabic question answering: A comparative study,” *arXiv preprint arXiv:2111.05671*, 2021.

[21] T. H. Alwaneen, A. M. Azmi, H. A. Aboalsamh, E. Cambria, and A. Hussain, “Arabic question answering system: A survey,” *Artificial Intelligence Review*, pp. 1–47, 2021.

[22] M. Aly and A. Atiya, “Labr: A large scale Arabic book reviews dataset,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 494–498.

[23] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based model for Arabic language understanding,” in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, 2020, p. 9.

[24] ——, “AraELECTRA: Pre-training text discriminators for Arabic language understanding,” in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, 2021, pp. 191–195.

[25] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, “Learning to retrieve reasoning paths over wikipedia graph for question answering,” *arXiv preprint arXiv:1911.10470*, 2019.

[26] Q. Ashur, *Su’al Wa Jawab Fi ALKORAN*. Beirut: Dar Ibn Hazm, 2001.

[27] A. Atef, B. Mattar, S. Sherif, E. Elrefai, and M. Torki, “AQAD: 17,000+ Arabic questions for machine comprehension of text,” in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2020, pp. 1–6.

[28] E. Atwell, N. Habash, B. Louw, *et al.*, “Understanding the quran: A new grand challenge for computer science and artificial intelligence,” *ACM-BCS Visions of Computer Science 2010*, 2010.

[29] M. Al-Azami, *The History of the Qur’anic Text from Revelation to Compilation: A Comparative Study with the Old and New Testaments*. London: UK Islamic Academy, 2003.

[30] ——, *The History of the Qur'anic Text from Revelation to Compilation: A Comparative Study with the Old and New Testaments, 2nd ed.* UK: Turath Publishing, 2020.

[31] A. M. Azmi and N. A. Alshenaifi, "LEMAZA: An Arabic why-question answering system," *Natural Language Engineering*, vol. 23, no. 6, pp. 877–903, 2017.

[32] W. Bakari and M. Neji, "A novel semantic and logical-based approach integrating RTE technique in the Arabic question–answering," *International Journal of Speech Technology*, pp. 1–17, 2020.

[33] W. Bakari, O. Trigui, and M. Neji, "Logic-based approach for improving Arabic question answering," in *2014 IEEE international conference on computational intelligence and computing research*, IEEE, 2014, pp. 1–6.

[34] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *Natural Language Engineering*, pp. 1–50, 2020.

[35] M. H. Bashir, A. M. Azmi, H. Nawaz, *et al.*, "Arabic natural language processing for Qur'anic research: A systematic review," *Artificial Intelligence Review*, pp. 1–54, 2022.

[36] F. Beirade, H. Azzoune, and D. E. Zegour, "Semantic query for quranic ontology," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 753–760, 2021.

[37] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[38] D. Chen, “Neural reading comprehension and beyond,” Ph.D. dissertation, Stanford University, 2018.

[39] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading wikipedia to answer open-domain questions,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2017, pp. 1870–1879.

[40] C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Jul. 2018, pp. 845–855.

[41] ——, “Simple and effective multi-paragraph reading comprehension,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Jul. 2018, pp. 845–855.

[42] J. H. Clark, E. Choi, M. Collins, *et al.*, “Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, 2020.

[43] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in *International Conference on Learning Representations*, 2020, p. 18.

[44] P. Cui, D. Hu, and L. Hu, “Listreader: Extracting list-form answers for opinion questions,” *arXiv preprint arXiv:2110.11692*, 2021.

[45] H. T. Dang, D. Kelly, and J. Lin, “Overview of the trec 2007 question answering track,” in *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2007)*, 2007.

[46] H. T. Dang, J. Lin, and D. Kelly, “Overview of the trec 2006 question answering track,” in *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2006)*, 2006.

[47] P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner, “QUOREF: A reading comprehension dataset with questions requiring coreferential reasoning,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5925–5932.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[49] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, “Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Jun. 2019, pp. 2368–2378.

[50] M. K. Eddine, N. Tomeh, N. Habash, J. L. Roux, and M. Vazirgiannis, “AraBART: A pretrained Arabic sequence-to-sequence model for abstractive summarization,” *arXiv:2203.10945 [cs]*, Mar. 2022.

[51] M. ElKomy and A. M. Sarhan, “TCE at Qur’an QA 2022: Arabic Language Question Answering Over Holy Qur’an Using a Post-Processed Ensemble of BERT-based Models,” in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[52] A. M. Ezzeldin, M. H. Kholief, and Y. El-Sonbaty, “Alqasim: Arabic language question answer selection in machines,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2013, pp. 100–103.

[53] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 3929–3938.

[54] M. Habash, *Mushaf Al Tajweed*. Syria: Dar-Al-Maarifa, 2001.

[55] A. Hakkoum and S. Raghay, “Semantic q&a system on the qur’an,” *Arabian Journal for Science and Engineering*, vol. 41, no. 12, pp. 5205–5214, Dec. 2016, ISSN: 1319-8025, 2191-4281.

[56] ——, “Semantic q&a system on the quran,” en, *Arabian Journal for Science and Engineering*, vol. 41, no. 12, pp. 5205–5214, Dec. 2016, ISSN: 1319-8025, 2191-4281.

[57] M. Hamdelsayed and E. Atwell, “Islamic applications of automatic question-answering,” *Journal of Engineering and Computer Science*, vol. 17, no. 2, pp. 51–57, 2016.

[58] M. A. Hamdelsayed and E. Atwell, “Using Arabic numbers (singular, dual, and plurals) patterns to enhance question answering system results,” in *IMAN’2016 4th International Conference on Islamic Applications in Computer Science and Technologies*, Leeds, 2016.

[59] M. A. Hamdelsayed, E. M. E. Mohamed, M. T. M. Saeed, *et al.*, “Islamic application of question answering systems: Comparative study,” *Journal of Advanced Computer Science and Technology Research*, vol. 7, no. 1, pp. 29–41, 2017.

[60] B. Hammo, A. Sleit, and M. El-Haj, “Effectiveness of query expansion in searching the holy quran,” 2007.

[61] B. Hamoud and E. Atwell, “Using an islamic question and answer knowledge base to answer questions about the holy quran,” *International Journal on Islamic Applications in Computer Science And Technology*, vol. 4, no. 4, pp. 20–29, 2016.

[62] ——, “Evaluation corpus for restricted-domain question-answering systems for the holy quran,” *International Journal of Science and Research*, vol. 6, no. 8, pp. 1133–1138, 2017.

[63] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” in *International Conference on Learning Representations*, 2021.

[64] M. Hu, Y. Peng, Z. Huang, and D. Li, “A multi-type multi-span network for reading comprehension that requires discrete reasoning,” in *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 1596–1606.

[65] W. S. Ismail and M. N. Homsi, "Dawqas: A dataset for Arabic why question answering system," *Procedia Computer Science*, Arabic Computational Linguistics, vol. 142, pp. 123–131, Jan. 2018, ISSN: 1877-0509.

[66] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[67] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2017, pp. 1601–1611.

[68] V. Karpukhin, B. Oguz, S. Min, *et al.*, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781.

[69] A. Keleg and W. Magdy, "Smash at qur'an qa 2022: Creating better faithful data splits for low-resourced question answering scenarios," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[70] I. A. El-Khair, "1.5 billion words Arabic corpus," *arXiv preprint arXiv:1611.04033*, 2016.

[71] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 252–262.

[72] O. Khattab, C. Potts, and M. Zaharia, "Relevance-guided supervision for OpenQA with ColBERT," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 929–944, 2021.

[73] K. Kishida, *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

[74] T. Kočiský, J. Schwarz, P. Blunsom, *et al.*, "The narrativeqa reading comprehension challenge," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 317–328, 2018.

[75] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Sep. 2017, pp. 785–794.

[76] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[77] ——, “The measurement of observer agreement for categorical data,” *Biometrics*, pp. 159–174, 1977.

[78] M. Lewis, Y. Liu, N. Goyal, *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[79] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, “MLQA: Evaluating cross-lingual extractive question answering,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 7315–7330.

[80] J. Lin and B. Katz, “Building a reusable test collection for question answering,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 851–861, 2006.

[81] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira, “Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations,” *arXiv preprint arXiv:2102.10073*, 2021.

[82] J. Lin, R. Nogueira, and A. Yates, “Pretrained transformers for text ranking: Bert and beyond,” *Synthesis Lectures on Human Language Technologies*, vol. 14, no. 4, pp. 1–325, 2021.

[83] Y. Liu, M. Ott, N. Goyal, *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” in *International Conference on Learning Representations*, 2020.

[84] H. M. Makhlouf, *Kalimat Al-Qur'an*. Beirut: Dar Ibn Hazm, 1997.

[85] R. Malhas and T. Elsayed, "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*, vol. 19, no. 6, pp. 1–21, Nov. 2020.

[86] ——, "Arabic Machine Reading Comprehension on the Holy Qur'an using CL-AraBERT," *Information Processing & Management*, vol. 59, no. 6, Nov. 2022.

[87] R. Malhas, W. Mansour, and T. Elsayed, "Qur'an QA 2022: Overview of the first shared task on question answering over the Holy Qur'an," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022, pp. 79–87.

[88] R. Malhas, M. Torki, R. Ali, T. Elsayed, and E. Yulianti, "Real, live, and concise: Answering open-domain questions with word embedding and summarization.," in *TREC*, 2016.

[89] R. Malhas, M. Torki, and T. Elsayed, "QU-IR at SemEval 2016 Task 3: Learning to rank on Arabic community question answering forums with word embedding," Association for Computational Linguistics (ACL), 2016.

[90] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[91] Y. Mellah, I. Touahri, Z. Kaddari, Z. Haja, J. Berrich, and T. Bouchentouf, "LARSA22 at Qur'an QA 2022: Text-to-Text Transformer for Finding Answers

to Questions from Qur'an," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[92] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[93] S. Min, D. Chen, L. Zettlemoyer, and H. Hajishirzi, "Knowledge guided text retrieval and reading for open domain question answering," *arXiv preprint arXiv:1911.03868*, 2019.

[94] S. Min, V. Zhong, R. Socher, and C. Xiong, "Efficient and robust question answering from minimal context over documents," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1725–1735.

[95] E. H. Mohamed and E. M. Shokry, "Qsst: A quranic semantic search tool based on word embedding," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[96] A. Mostafa and O. Mohamed, "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 *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[97] H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic question answering," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Association for Computational Linguistics, Aug. 2019, pp. 108–118.

[98] D. L. Newman, "The Arabic literary language: The nahda and beyond," *The Oxford Handbook of Arabic Linguistics*, p. 472, 2013.

[99] Y. Nie, S. Wang, and M. Bansal, "Revealing the importance of semantic retrieval for machine reading at scale," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Nov. 2019, pp. 2553–2566.

[100] K. Nishida, I. Saito, A. Otsuka, H. Asano, and J. Tomita, "Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 647–656.

[101] R. Nogueira and K. Cho, "Passage re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.

[102] K. Ouda, "Qurananalysis: A semantic search and intelligence system for the quran," Ph.D. dissertation, Master Thesis, University of Leeds, Leeds, UK, 2015.

[103] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback, 2022," *URL https://arxiv. org/abs/2203.02155*,

[104] A. Peñas, E. Hovy, P. Forner, A. Rodrigo, R. Sutcliffe, and R. Morante, "Qa4mre 2011-2013: Overview of question answering for machine reading evaluation,"

in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2013, pp. 303–320.

[105] A. Peñas, E. H. Hovy, P. Forner, *et al.*, "Overview of qa4mre at clef 2011: Question answering for machine reading evaluation.," in *CLEF (Notebook Papers/Labs/Workshop)*, 2012, pp. 1–20.

[106] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations.," in *NAACL*, Association for Computational Linguistics, 2018.

[107] F. Petroni, T. Rocktäschel, S. Riedel, *et al.*, "Language models as knowledge bases?" In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473.

[108] D. Premasiri, T. Ranasinghe, W. Zaghouani, R. Mitkov, J. Berrich, and T. Bouchentouf, "DTW at Qur'an QA 2022: Utilising Transfer Learning with Transformers for Question Answering in a Low-resource Domain," in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[109] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Technical report, OpenAI*, 2018.

[110] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the As-*

sociation for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789.

[111] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2016, pp. 2383–2392.

[112] M. Richardson, C. J. Burges, and E. Renshaw, “MCTest: A challenge dataset for the open-domain machine comprehension of text,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 193–203.

[113] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, “Okapi at trec-3,” in *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, 1994, pp. 109–126.

[114] M. Romanov and M. Seydi, “Openiti: A machine-readable corpus of islamicate texts,” Zenodo, May 2019. [Online]. Available: `https://zenodo.org/record/3082464`.

[115] M. K. Saad and W. M. Ashour, “Osac: Open source Arabic corpora,” in *6th ArchEng Int. Symposiums, EEECS*, vol. 10, 2010.

[116] A. Safaya, M. Abdullatif, and D. Yuret, “Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 2054–2059.

[117] M. A. M. Safee, M. M. Saudi, S. A. Pitchay, *et al.*, “Hybrid search approach for retrieving medical and health science knowledge from quran,” *International Journal of Engineering and Technology (UAE)*, 2018.

[118] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018.

[119] E. Segal, A. Efrat, M. Shoham, A. Globerson, and J. Berant, “A simple and effective model for answering multi-span questions,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3074–3080.

[120] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” *arXiv preprint arXiv:1611.01603*, 2016.

[121] A.-B. M. Sharaf and E. Atwell, “Qurana: Corpus of the quran annotated with pronominal anaphora.,” in *LREC*, Citeseer, 2012, pp. 130–137.

[122] M. A. Sherif and A.-C. N. Ngomo, “Semantic quran,” *Semantic Web*, vol. 6, no. 4, pp. 339–345, 2015.

[123] H. Shmeisani, S. Tartir, A. Al-Na’ssaan, and M. Naji, “Semantically answering questions from the holy quran,” in *International Conference on Islamic Applications in Computer Science And Technology*, 2014, pp. 1–8.

[124] J. Sim and C. C. Wright, “The kappa statistic in reliability studies: Use, interpretation, and sample size requirements,” *Physical Therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[125] ——, “The kappa statistic in reliability studies: Use, interpretation, and sample size requirements,” *Physical Therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[126] N. Singh, “niksss at Qur’an QA 2022: A Heavily Optimized BERT Based Model for Answering Questions from the Holy Qu’ran,” in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[127] M. N. Swar, *Mushaf Al-Tafseel Al-Mawdoo’ee*. Damascus: Dar Al-Fajr Al-Islami, 2007.

[128] I. J. AT-Tabari, *Tafsir Ibn Jarir AT-Tabari - Jami’ Al-Bayan ’an Ta-wil Al-Quran (923)*. Beirut: Dar al-Kutub al-’Ilmiyah. (Original work published 923), 1997.

[129] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[130] E. M. Voorhees, “Overview of the trec 2003 question answering track,” in *Proceedings of the Eleventh Text REtrieval Conference (TREC 2003)*, 2003.

[131] E. M. Voorhees and H. T. Dang, “Overview of the trec 2005 question answering track,” in *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2005)*, 2005, pp. 52–62.

[132] E. M. Voorhees and D. M. Tice, “Building a question answering test collection,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2000, pp. 200–207.

[133] E. M. Voorhees, “Overview of the trec 2004 question answering track,” in *Proceedings of the Twelfth Text REtrieval Conference (TREC 2004)*, 2004, pp. 54–68.

[134] B. Wang, S. Guo, K. Liu, S. He, and J. Zhao, “Employing external rich knowledge for machine comprehension.,” in *IJCAI*, 2016, pp. 2929–2935.

[135] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, “Multi-passage BERT: A globally normalized BERT model for open-domain question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5878–5882.

[136] A. Wasfey, E. Elrefai, M. Marwa, and N. Haq, “Stars at Qur’an QA 2022: Building Automatic Extractive Question Answering Systems for the Holy Qur’an with Transformer Models and Releasing a New Dataset,” in *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, 2022.

[137] Y. Wu, M. Schuster, Z. Chen, *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.

[138] L. Xue, N. Constant, A. Roberts, *et al.*, “Mt5: A massively multilingual pre-trained text-to-text transformer,” in *NAACL-HLT*, 2021.

[139] J. Yang, Z. Zhang, and H. Zhao, “Multi-span style extraction for generative reading comprehension,” *arXiv preprint arXiv:2009.07382*, 2020.

[140] W. Yang, Y. Xie, A. Lin, *et al.*, “End-to-end open-domain question answering with BERTserini,” in *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Jun. 2019, pp. 72–77.

[141] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[142] Z. Yang, P. Qi, S. Zhang, *et al.*, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2369–2380.

[143] A. R. Yauri, R. A. Kadir, A. Azman, and M. A. Murad, "Quranic verse extraction base on concepts using owl-dl ontology," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 23, pp. 4492–4498, 2013.

[144] A. W. Yu, D. Dohan, M.-T. Luong, *et al.*, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[145] N. Yusuf, M. A. M. Yunus, N. Wahid, N. M. Nawi, N. A. Samsudin, and N. Arbaiy, "Query expansion method for quran search using semantic search and lucene ranking," *J Eng Sci Technol*, vol. 15, no. 1, pp. 675–692, 2020.

[146] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets," *Applied Sciences*, vol. 10, no. 2121, p. 7640, Jan. 2020.

[147] I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja, "Osian: Open source international Arabic news corpus-preparation and integration into the clarin-

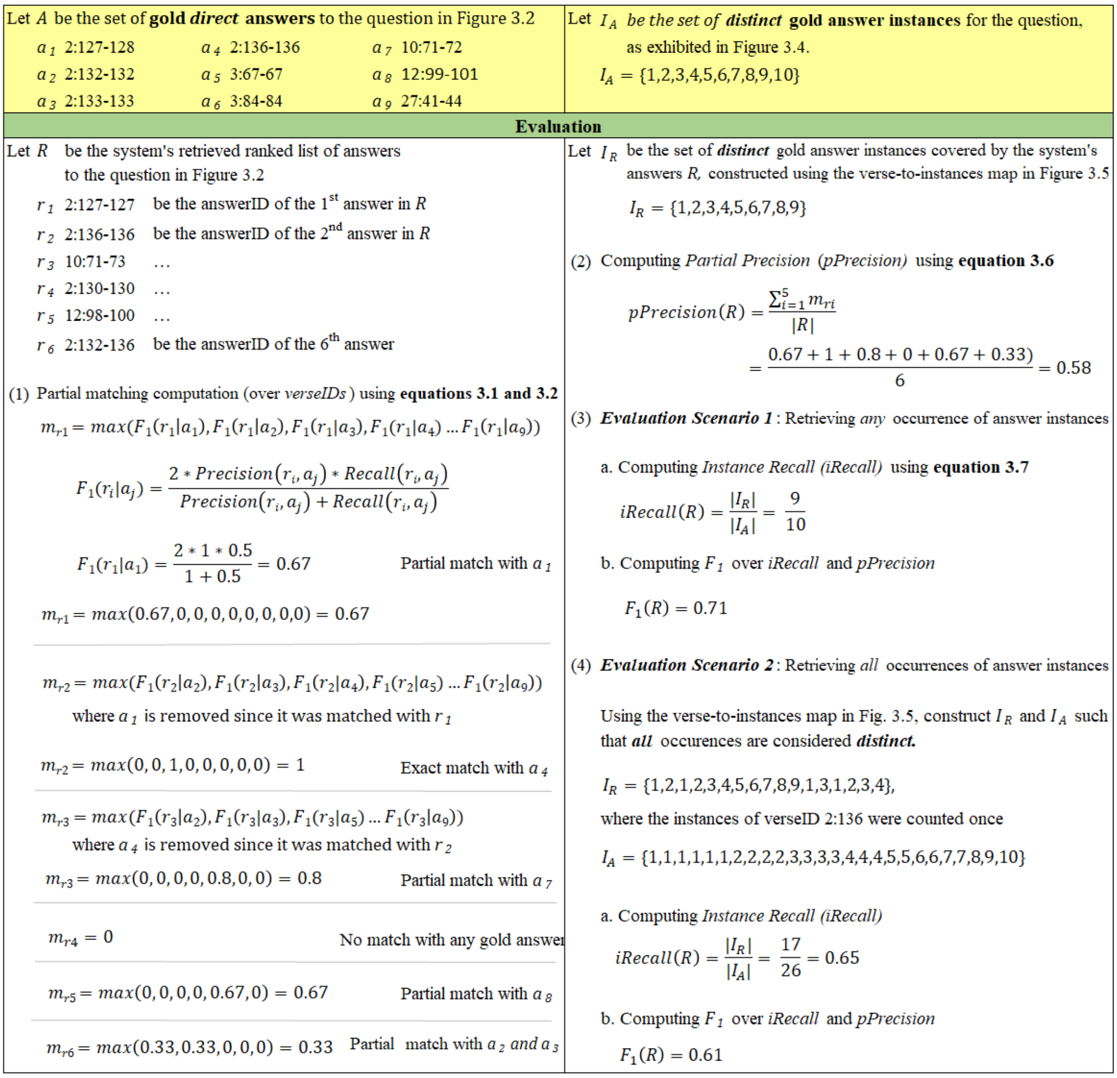
infrastructure," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 175–182.

[148] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," *arXiv:2101.00774 [cs]*, May 2021.

[149] S. Zouaoui and K. Rezeg, "A novel quranic search engine using an ontology-based semantic indexing," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3653–3674, 2021.

APPENDIX A: EVALUATION EXAMPLE FOR VERSE-BASED ANSWERS

In this appendix, we present a full example for evaluating a QA system that returns answers in terms of Qur'anic verses. Figure A.1 showcases the evaluation of a system on a multi-answer question assuming two evaluation scenarios, as explained in Section 3.1.6. The first scenario evaluates the system for retrieving *any* occurrence of answer instances (step (3) in Figure A.1), while the second evaluates the system for retrieving *all* occurrences of answer instances (step (4) in Figure A.1).



Let A be the set of **gold *direct* answers** to the question in Figure 3.2

a_1	2:127-128	a_4	2:136-136	a_7	10:71-72
a_2	2:132-132	a_5	3:67-67	a_8	12:99-101
a_3	2:133-133	a_6	3:84-84	a_9	27:41-44

Let I_A *be the set of **distinct*** **gold answer instances** for the question, as exhibited in Figure 3.4.

$I_A = \{1,2,3,4,5,6,7,8,9,10\}$

Evaluation

Let R be the system's retrieved ranked list of answers to the question in Figure 3.2

r_1	2:127-127	be the answerID of the 1st answer in R
r_2	2:136-136	be the answerID of the 2nd answer in R
r_3	10:71-73	…
r_4	2:130-130	…
r_5	12:98-100	…
r_6	2:132-136	be the answerID of the 6th answer

(1) Partial matching computation (over *verseIDs*) using **equations 3.1 and 3.2**

$$m_{r1} = max(F_1(r_1|a_1), F_1(r_1|a_2), F_1(r_1|a_3), F_1(r_1|a_4) \ldots F_1(r_1|a_9))$$

$$F_1(r_i|a_j) = \frac{2 * Precision(r_i, a_j) * Recall(r_i, a_j)}{Precision(r_i, a_j) + Recall(r_i, a_j)}$$

$$F_1(r_1|a_1) = \frac{2 * 1 * 0.5}{1 + 0.5} = 0.67$$ Partial match with a_1

$$m_{r1} = max(0.67, 0, 0, 0, 0, 0, 0, 0, 0) = 0.67$$

$$m_{r2} = max(F_1(r_2|a_2), F_1(r_2|a_3), F_1(r_2|a_4), F_1(r_2|a_5) \ldots F_1(r_2|a_9))$$

where a_1 is removed since it was matched with r_1

$$m_{r2} = max(0, 0, 1, 0, 0, 0, 0, 0) = 1$$ Exact match with a_4

$$m_{r3} = max(F_1(r_3|a_2), F_1(r_3|a_3), F_1(r_3|a_5) \ldots F_1(r_3|a_9))$$

where a_4 is removed since it was matched with r_2

$$m_{r3} = max(0, 0, 0, 0, 0.8, 0, 0) = 0.8$$ Partial match with a_7

$$m_{r4} = 0$$ No match with any gold answer

$$m_{r5} = max(0, 0, 0, 0, 0.67, 0) = 0.67$$ Partial match with a_8

$$m_{r6} = max(0.33, 0.33, 0, 0, 0) = 0.33$$ Partial match with a_2 *and* a_3

Let I_R be the set of ***distinct*** gold answer instances covered by the system's answers R, constructed using the verse-to-instances map in Figure 3.5

$I_R = \{1,2,3,4,5,6,7,8,9\}$

(2) Computing *Partial Precision (pPrecision)* using **equation 3.6**

$$pPrecision(R) = \frac{\sum_{i=1}^{5} m_{ri}}{|R|}$$

$$= \frac{0.67 + 1 + 0.8 + 0 + 0.67 + 0.33)}{6} = 0.58$$

(3) ***Evaluation Scenario 1***: Retrieving *any* occurrence of answer instances

a. Computing *Instance Recall (iRecall)* using **equation 3.7**

$$iRecall(R) = \frac{|I_R|}{|I_A|} = \frac{9}{10}$$

b. Computing F_1 over *iRecall* and *pPrecision*

$$F_1(R) = 0.71$$

(4) ***Evaluation Scenario 2***: Retrieving *all* occurrences of answer instances

Using the verse-to-instances map in Fig. 3.5, construct I_R and I_A such that ***all*** occurences are considered ***distinct.***

$I_R = \{1,2,1,2,3,4,5,6,7,8,9,1,3,1,2,3,4\}$,

where the instances of verseID 2:136 were counted once

$I_A = \{1,1,1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,5,5,6,6,7,7,8,9,10\}$

a. Computing *Instance Recall (iRecall)*

$$iRecall(R) = \frac{|I_R|}{|I_A|} = \frac{17}{26} = 0.65$$

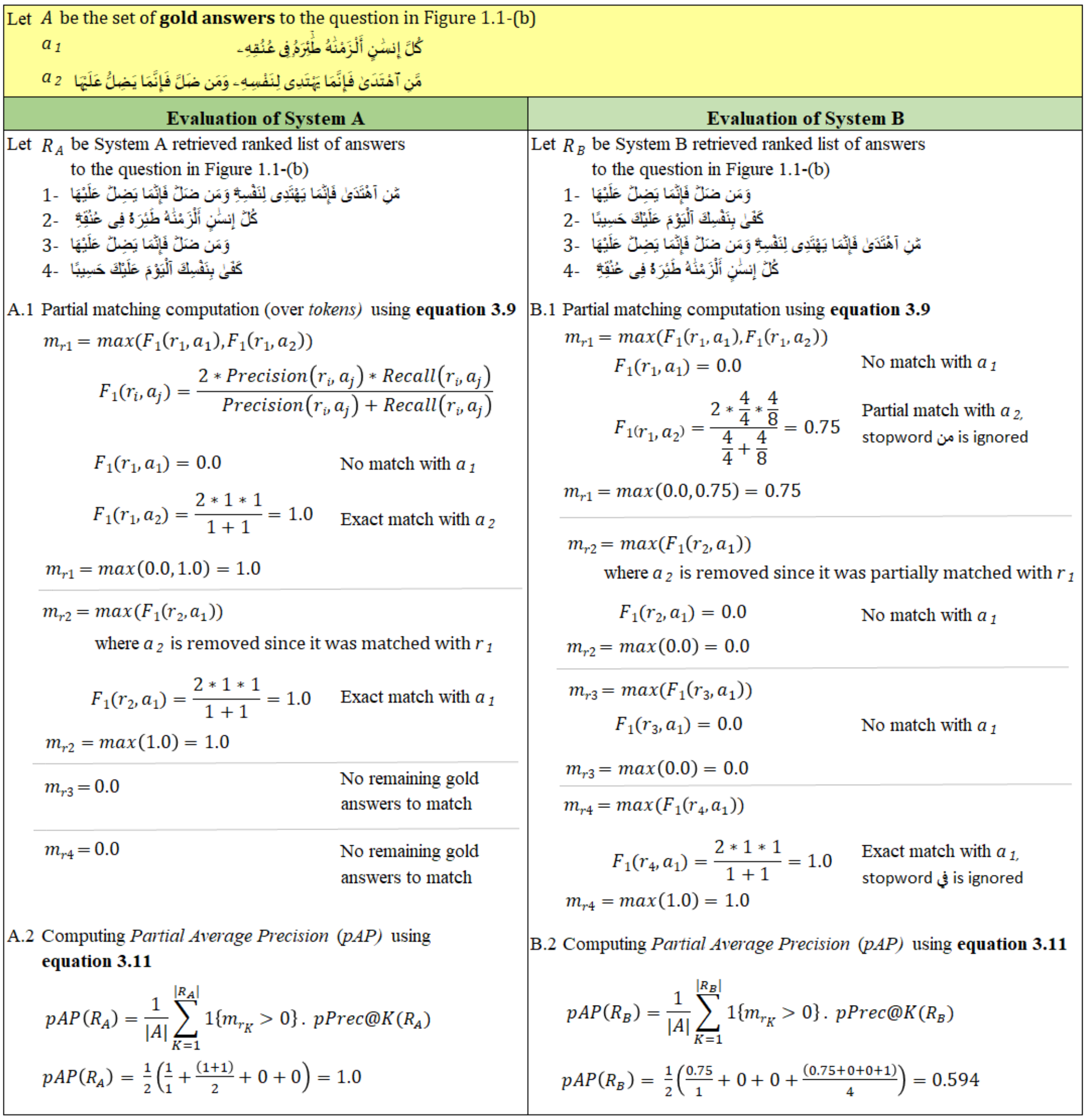
b. Computing F_1 over *iRecall* and *pPrecision*

$$F_1(R) = 0.61$$

Figure A.1. Evaluation example of a system's response to a multi-answer question. The answers are verse-based.

APPENDIX B: EVALUATION EXAMPLE FOR SPAN-BASED ANSWERS

In this appendix, we present a full example for evaluating extractive QA/MRC systems (that return answers as spans of text). Figure B.1 shows how the proposed rank-based measure (Partial Average Precision) pAP is computed, as explained in Section 3.2.3. The example compares the performance of two different systems given the same multi-answer question, to showcase its fairness. System A attains a better pAP score than system B although both predict the same set of answers *but* in different ordering. pAP perfectly rewards system A since it *exactly* predicts the two correct answers at ranks 1 and 2, while system B predicts the first correct answer *partially* at rank 1, and predicts the second answer *exactly* at rank 4.



Let A be the set of **gold answers** to the question in Figure 1.1-(b)

a_1 كُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ

a_2 مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ ۖ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا

Evaluation of System A

Let R_A be System A retrieved ranked list of answers to the question in Figure 1.1-(b)

1- مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا
2- كُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ
3- وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا
4- كَفَىٰ بِنَفْسِكَ ٱلْيَوْمَ عَلَيْكَ حَسِيبًا

A.1 Partial matching computation (over *tokens*) using **equation 3.9**

$$m_{r1} = max(F_1(r_1, a_1), F_1(r_1, a_2))$$

$$F_1(r_i, a_j) = \frac{2 * Precision(r_i, a_j) * Recall(r_i, a_j)}{Precision(r_i, a_j) + Recall(r_i, a_j)}$$

$F_1(r_1, a_1) = 0.0$ No match with a_1

$F_1(r_1, a_2) = \frac{2 * 1 * 1}{1 + 1} = 1.0$ Exact match with a_2

$m_{r1} = max(0.0, 1.0) = 1.0$

$m_{r2} = max(F_1(r_2, a_1))$

where a_2 is removed since it was matched with r_1

$F_1(r_2, a_1) = \frac{2 * 1 * 1}{1 + 1} = 1.0$ Exact match with a_1

$m_{r2} = max(1.0) = 1.0$

$m_{r3} = 0.0$ No remaining gold answers to match

$m_{r4} = 0.0$ No remaining gold answers to match

A.2 Computing *Partial Average Precision* (*pAP*) using **equation 3.11**

$$pAP(R_A) = \frac{1}{|A|} \sum_{K=1}^{|R_A|} 1\{m_{r_K} > 0\}\,.\ pPrec@K(R_A)$$

$$pAP(R_A) = \frac{1}{2}\left(\frac{1}{1} + \frac{(1+1)}{2} + 0 + 0\right) = 1.0$$

Evaluation of System B

Let R_B be System B retrieved ranked list of answers to the question in Figure 1.1-(b)

1- وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا
2- كَفَىٰ بِنَفْسِكَ ٱلْيَوْمَ عَلَيْكَ حَسِيبًا
3- مَّنِ ٱهْتَدَىٰ فَإِنَّمَا يَهْتَدِى لِنَفْسِهِۦ وَمَن ضَلَّ فَإِنَّمَا يَضِلُّ عَلَيْهَا
4- كُلَّ إِنسَٰنٍ أَلْزَمْنَٰهُ طَٰٓئِرَهُۥ فِى عُنُقِهِۦ

B.1 Partial matching computation using **equation 3.9**

$$m_{r1} = max(F_1(r_1, a_1), F_1(r_1, a_2))$$

$F_1(r_1, a_1) = 0.0$ No match with a_1

$F_1(r_1, a_2) = \frac{2 * \frac{4}{4} * \frac{4}{8}}{\frac{4}{4} + \frac{4}{8}} = 0.75$ Partial match with a_2, stopword من is ignored

$m_{r1} = max(0.0, 0.75) = 0.75$

$m_{r2} = max(F_1(r_2, a_1))$

where a_2 is removed since it was partially matched with r_1

$F_1(r_2, a_1) = 0.0$ No match with a_1

$m_{r2} = max(0.0) = 0.0$

$m_{r3} = max(F_1(r_3, a_1))$

$F_1(r_3, a_1) = 0.0$ No match with a_1

$m_{r3} = max(0.0) = 0.0$

$m_{r4} = max(F_1(r_4, a_1))$

$F_1(r_4, a_1) = \frac{2 * 1 * 1}{1 + 1} = 1.0$ Exact match with a_1, stopword في is ignored

$m_{r4} = max(1.0) = 1.0$

B.2 Computing *Partial Average Precision* (*pAP*) using **equation 3.11**

$$pAP(R_B) = \frac{1}{|A|} \sum_{K=1}^{|R_B|} 1\{m_{r_K} > 0\}\,.\ pPrec@K(R_B)$$

$$pAP(R_B) = \frac{1}{2}\left(\frac{0.75}{1} + 0 + 0 + \frac{(0.75+0+0+1)}{4}\right) = 0.594$$

Figure B.1. Full example of how pAP evaluation measure is computed given the returned answers of two different systems on the same multi-answer question. The answers are span-based.