Information Processing and Management 59 (2022) 103068

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Arabic machine reading comprehension on the Holy Qur'an using CL-AraBERT

Rana Malhas [*], Tamer Elsayed

Computer Science and Engineering Department, Qatar University, Qatar

ARTICLE INFO

Dataset link: https://github.com/RanaMalhas/QRCD

Keywords:
Classical Arabic
Reading comprehension
Answer extraction
Partial matching evaluation
Pre-trained language models
Cross-lingual transfer learning

ABSTRACT

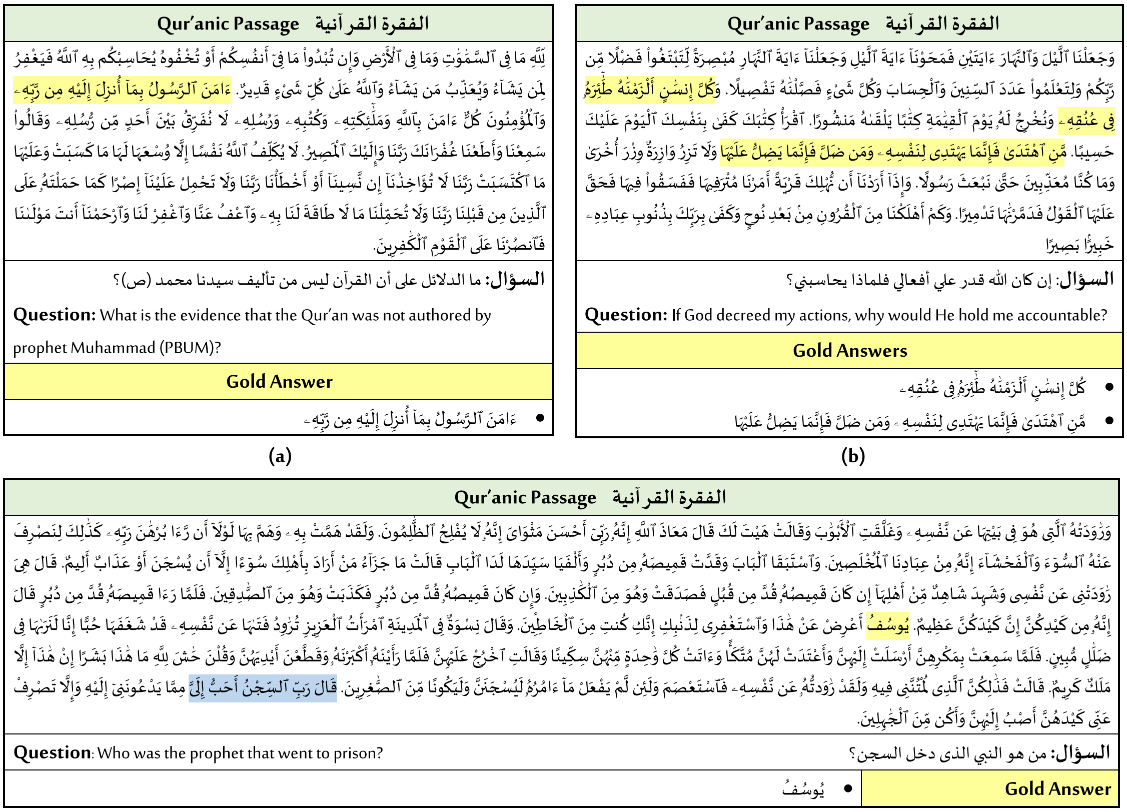
In this work, we tackle the problem of machine reading comprehension (MRC) on the Holy Qur'an to address the lack of Arabic datasets and systems for this important task. We construct *QRCD* as the first Qur'anic Reading Comprehension Dataset, composed of 1,337 question-passage-answer triplets for 1,093 question-passage pairs, of which 14% are multi-answer questions. We then introduce CLassical-AraBERT (CL-AraBERT for short), a new AraBERT-based pre-trained model, which is further pre-trained on about 1.0B-word Classical Arabic (CA) dataset, to complement the Modern Standard Arabic (MSA) resources used in pre-training the initial model, and make it a better fit for the task. Finally, we leverage cross-lingual transfer learning from MSA to CA, and fine-tune CL-AraBERT as a reader using two MSA-based MRC datasets followed by our *QRCD* dataset to constitute the first (to the best of our knowledge) MRC system on the Holy Qur'an. To evaluate our system, we introduce *Partial Average Precision* (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 MSA questions. Adopting two experimental evaluation setups (hold-out and cross validation (CV)), we empirically show that the fine-tuned CL-AraBERT reader model significantly outperforms the baseline fine-tuned AraBERT reader model by 6.12 and 3.75 points in pAP scores, in the hold-out and CV setups, respectively. To promote further research on this task and other related tasks on Qur'an and Classical Arabic text, we make both the *QRCD* dataset and the pre-trained CL-AraBERT model publicly available.

1. Introduction

Since its inception in the 1970s, reading comprehension was perceived as the ideal apparatus to evaluate the task of language understanding by computer systems (Chen, 2018). Given a passage of text, a machine reading comprehension (MRC) system should read this passage and answer comprehension questions about it Chen (2018). After being dormant for decades, the MRC field witnessed a resurgence that was mainly attributed to the development of large reading comprehension datasets (Joshi, Choi, Weld, & Zettlemoyer, 2017; Lai, Xie, Liu, Yang, & Hovy, 2017; Rajpurkar, Zhang, Lopyrev, & Liang, 2016), 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 (Chen, 2018). Moreover, the advent and phenomenal success of transformer-based pre-trained language models, e.g., BERT (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (Liu et al., 2020) and XLNet (Yang et al., 2019), have further escalated the rate at which the field of neural MRC was progressing.

[*] Corresponding author.
E-mail addresses: rana.malhas@qu.edu.qa (R. Malhas), telsayed@qu.edu.qa (T. Elsayed).

https://doi.org/10.1016/j.ipm.2022.103068
Received 26 February 2022; Received in revised form 1 August 2022; Accepted 18 August 2022
Available online 9 September 2022



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)

Fig. 1. Example MRC questions and answers. (a) A non-factoid single-answer question with an evidence-based answer that is a single span of text. (b) A non-factoid multi-answer 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 single-answer 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

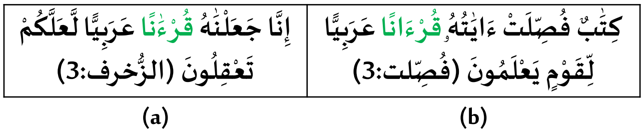
Interestingly, the perception towards the task has evolved from being a question answering (QA) task over a closed piece of text into an integral component of modern AI systems, e.g., the Retriever-Reader open-domain QA (Chen, Fisch, Weston, & Bordes, 2017; Clark & Gardner, 2018; Yang et al., 2019; Zhu et al., 2021) and conversational QA (Choi et al., 2018; Yatskar, 2019). 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 (Peñas et al., 2013) or answer generation (Baradaran, Ghiasi, & Amirkhani, 2020; Kočiský et al., 2018).

Among the main challenges hindering the progress of *Arabic* MRC (in comparison to English MRC) is the scarcity of large MRC datasets in Arabic, except for a few large and moderately sized ones, e.g., the Arabic SQuAD and ARCD (Mozannar, Maamary, El Hajal, & Hajj, 2019), and the DAWQAS dataset (Ismail & Homsi, 2018). Interestingly, all those MRC datasets are in Modern Standard Arabic (MSA). To the best of our knowledge of the literature, there are neither *extractive*[1] MRC datasets nor systems that tackle the *Qur'anic Classical Arabic* (CA), in which the Holy Qur'an is written (though there are several search tools and QA systems). Revealed fourteen centuries ago, the Qur'an is sacredly held by more than 1.8B Muslims across the world.[2] It is composed of 114 chapters (Suras) and 6236 verses that comprise more than 80k words in Classical Arabic (Bashir et al., 2021). Being the main source of teachings, knowledge, and legislation in Islam, the Holy Qur'an is sought and explored by Muslims and non-Muslims, looking for answers to their questions out of religiosity, curiosity, or skepticism (Malhas & Elsayed, 2020). With the recent resurgence of the MRC field and the permanent interest in Qur'an, there is a clear need for building MRC systems for the Qur'an.

In this work, we address the problem of MRC on the Holy Qur'an. The problem is defined as follows: given a Qur'anic passage p and a question posed in MSA q, an extractive MRC system S should extract and return the answer(s) R to the given question

[1] *Extractive* MRC refers to the task of span prediction, where the answer is a specific span of text that is *extracted* (rather than *generated*) from a passage accompanying a question (Baradaran et al., 2020; Zhu et al., 2021).

[2] https://en.wikipedia.org/wiki/Islam_by_country.



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

Fig. 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

q from the accompanying passage p. An answer r is a span of text extracted from the passage p, denoted as an "answer span". Fig. 1 illustrates the task of MRC on the Holy Qur'an using three different examples of question-passage pairs along with their answers. Since a question might generally have multiple answers in the passage (such as the question in Fig. 1-(b)), the system should return a *list* of predicted answers; the list should be *ranked* so that the user (who asked the question) can intuitively check the top (potentially correct) answers before moving down the list. Therefore, the predicted answers R that the system S returns is expected to be a *ranked list* of answers. 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.[3] Broadly, questions are categorized into two abstract types: *single-answer* and *multi-answer* questions. A *single-answer* question has only one answer span as shown in Fig. 1-(a) and (c). Whereas, a *multi-answer* question may have two or more different answer spans (in distant or contiguous verses) in the accompanying Qur'anic passage. Each answer span represents an answer component that may answer the question (fully or partially), as shown in Fig. 1-(b). In general, due to the literary style of Qur'anic text, answers to non-factoid questions are mostly evidence-based (Fig. 1-(a) and (b)), while answers to factoid questions are likely to require some form of coreference resolution (Fig. 1-(c)). Consequently, our MRC task on the Qur'an requires multi-verse reasoning.[4]

Qur'anic CA text has its own set of challenges to Natural Language Processing (NLP) tasks (including MRC). The Holy Qur'an is a phenomenal collection that is characterized with its long-chained anaphoric structures across the verses of the same chapter (Fig. 1-(c)). Other than the titles of chapters, the Qur'an has no subtitles, which makes the topology of its diverse topics quite challenging. Each chapter may tackle several topics, and each verse may relate to more than one topic. Furthermore, the same topic may be addressed in many different chapters/verses, but in variant contexts (Malhas & Elsayed, 2020); we denote this feature by "unstructured topic diversity". Moreover, it is of paramount importance for the MRC task to address the challenge of bridging the gap between the questions being in MSA and the answers being in Qur'anic CA; we denote this gap by the "*MSA-to-CA gap*" (Fig. 1-(b) presents an example of such gap). The latter challenge is further compounded due to the rather sporadic non-conformity of the Holy Qur'an's Uthmani orthography[5] to Classical Arabic (as shown in Fig. 2), which is an open issue in Qur'anic NLP research (Bashir et al., 2021).

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 and obsolete words were dropped (Newman, 2013). Nevertheless, CA remains richer in lexis (Sharaf & Atwell, 2012), which widens the MSA-to-CA gap. Other general challenges inherent in the Arabic language include the absence of capital letters and lack of diacritics in MSA. Diacritics are important because they affect the meaning hence understanding of Arabic text. 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 (Bashir et al., 2021). In such cases, only the context of words is used to disambiguate their intended meaning, which poses clear challenges to NLP systems.

In this work, we address the lack of Arabic *resources* and *systems* for extractive MRC on the Holy Qur'an in the literature. We first introduce ***QRCD*** as the first **Q**ur'anic **R**eading **C**omprehension **D**ataset that adopts the same format of SQuAD v1.1 (Rajpurkar et al., 2016). *QRCD* is composed of 1337 question-passage-answer triplets for 1093 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 the MRC task. Second, we introduce **CL-AraBERT** (denoting CLassical AraBERT), a new AraBERT-based (Antoun, Baly, & Hajj, 2020) 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. Finally, 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. The goal is to bridge the MSA-to-CA gap, and mitigate the lack of large MRC datasets in CA, hence constituting the first (to the best of our knowledge) MRC system on the Holy Qur'an.

To evaluate our work, two experimental setups were used: a 75%–25% train-holdout setup and a 5-fold cross validation (CV) setup over the *QRCD* dataset. Our experiments show that the fine-tuned CL-AraBERT reader model significantly outperformed the baseline fine-tuned AraBERT model in the hold-out and CV setups, respectively.

Our contribution in this work is five-fold:

[3] 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.

[4] We organized Quran QA 2022 shared task to triggor MRC research on the Holy Qur'an (Malhas, Mansour, & Elsayed, 2022)

[5] 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 (Al-Azami, 2020; Bashir et al., 2021).

(1) We introduce *QRCD* as the first extractive machine reading comprehension dataset on the Holy Qur'an.
(2) We introduce CL-AraBERT, which is a further pre-trained version of the AraBERT model (Antoun et al., 2020), using a large Classical Arabic dataset. We then fine-tune it to constitute the first extractive MRC system on the Holy Qur'an.
(3) We introduce a simple yet novel method to fairly match predicted answers of multi-answer questions against their respective gold answers, and 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.
(4) 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.
(5) We make the pre-trained CL-AraBERT model, the *QRCD* training set, and the evaluation script publicly available to promote state-of-the-art research on the task.[6]

The rest of the article is organized as follows. Section 2 covers related work, which is followed by a description of our methodology in developing the *QRCD* dataset in Section 3. Section 4 introduces our methodology in developing the CL-AraBERT reader. Section 5 describes the evaluation measures we used. Section 6 presents and discusses our experimental evaluation results and their implications, in addition to a qualitative performance analysis using failure and success examples. We conclude this section with general implications of our research work. Then in Section 7, we conclude with final remarks and some future directions.

2. Related work

Contrary to the ubiquitous presence of English MRC datasets and systems in the literature, MRC datasets and systems in Arabic are scarce, and are only present in Modern Standard Arabic. To the best of our knowledge, there are no extractive MRC datasets or systems on the Holy Qur'an in the literature, though there are several QA datasets and systems on the Qur'an that we briefly overview for contrastive purposes (Section 2.1). Then, we overview existing Arabic reading comprehension datasets and systems (Section 2.2) before discussing important transformer-based MRC models in the literature (Section 2.3).

2.1. Existing QA datasets and systems on the Holy Qur'an

Despite the presence of some question answering datasets on the Holy Qur'an (Alqahtani & Atwell, 2018; Hamdelsayed & Atwell, 2016; Hamoud & Atwell, 2017; Malhas & Elsayed, 2020), they lack the fundamental property that characterizes reading comprehension datasets of coupling QA pairs with the corresponding passages or documents (i.e., contexts) to which the answers where extracted from. Hamoud and Atwell developed the QAEQ&AC (Qur'an Arabic-English Question and Answer Corpus), which is a QA dataset of 1500 questions, 1000 of which are in Arabic and the remaining are in English. The majority of the questions are natural questions (i.e., posed by real inquisitors) that were collected from variant sources with their answers being mainly in natural language text, in addition to some answers being Qur'anic verses. Hamdelsayed and Atwell developed a QA dataset of 263 question answer pairs; the questions and their answers were drawn from the first two chapters of the Holy Qur'an (Al-Fatiha and Al-Baqara). Alqahtani and Atwell developed the AQQAC (Annotated Corpus of Arabic Al-Qur'an Question and Answer) dataset. It is composed of 1224 QA pairs (in addition to 1000 unpublished ones due to copyright concerns) that have natural language answers generated from Tafseer Al-Tabari; each answer is accompanied by its respective verse-based answer. The QA pairs were scrapped from a website on Al-Qur'an and Tafseer. Unlike our *QRCD* dataset, which is an expansion of AyaTEC (Malhas & Elsayed, 2020), the aforementioned datasets (except for AQQAC) are not publicly available, and they did not include contexts to their question answer pairs. Moreover, except for AyaTEC, the answers to the questions do not cover the whole Qur'an.

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 literature review focused on the answer extraction components of Qur'anic QA systems (not search tools), to explore if any could be perceived as a reading comprehension component. Hamoud and Atwell (2016) developed a simple QA system over their QAEQ&AC corpus (Hamoud & Atwell, 2017) that was developed as a search engine over the questions of this corpus. A question posed in natural language is best matched to the top retrieved similar question, whose answer was returned as the answer to the posed question. Hakkoum and Raghay (2016) developed a semantic QA system by developing an ontology over the Qur'an knowledge and concepts, and a natural language interface that processes and transforms questions (posed in Arabic) into formal ontology queries, which are then used to retrieve the answers to the questions from the ontology. The semantic-based approach adopted by Shmeisani, Tartir, Al-Na'ssaan, and Naji (2014) for their QA system is highly similar to that of Hakkoum and Raghay (2016), but it was only applied on factoid questions. None of the previous QA systems have answer extraction components that resemble a reading comprehension component.

On the other hand, Abdelnasser et al. (2014) developed a semantic-based QA system (Al-Bayan) that accepts a question in Arabic, then retrieves the concept-matching verses to the question from a Qur'anic ontology (of 1217 concepts). This ontology classifies verses according to their topics/concepts; each leaf concept in the ontology is linked to its relevant verses and respective interpretations from two Tafseer books (Ibn-Kathir and Al-Jaza'iri). Finally, the system's answer extraction component, extracts (or generates) the answer from the retrieved Qur'anic verses and their interpretations using a Named Entity Recognition (NER)

[6] All can be downloaded from this link https://github.com/RanaMalhas/QRCD.

model. We believe that their semantic-based answer extraction methodology could have been considered an extractive reading comprehension component on the Qur'an, if the answers were only extracted from the Qur'anic text without its interpretations (i.e., the extracted answers need not be Qur'anic verses only). A limitation of this system is its design to answer factoid questions only. On the other hand, our proposed approach handles both factoid and non-factoid questions, and establishes the first extractive reading comprehension system on the Holy Qur'an.

2.2. 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 (Akour, Abufardeh, Magel, & Al-Radaideh, 2011) 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 organized at the CLEF (Cross-Language Evaluation Forum) for several languages with Arabic being one of them (Peñas et al., 2012). 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 (Abouenour, Bouzoubaa, & Rosso, 2012) and ALQASIM (Ezzeldin, Kholief, & El-Sonbaty, 2013) 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) (Mann & Thompson, 1988) was proposed by Azmi and Alshenaifi (2017) 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 (Alwaneen, Azmi, Aboalsamh, Cambria, & Hussain, 2021). LEMAZA was evaluated using 110 *why* questions over a dataset of 700 articles extracted from the OSAC Arabic corpus (Saad & Ashour, 2010). 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 (Alwaneen et al., 2021; Bakari & Neji, 2020; Bakari, Trigui, & Neji, 2014). Starting from 2018 onwards, relatively larger Arabic MRC datasets started to appear in the literature. Ismail and Homsi (2018) 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 et al. (2019). 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 1395 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 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 et al. (2017). It is composed of a TF-IDF document retriever and a fine-tuned multilingual BERT (Devlin et al., 2019) 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 (Antoun et al., 2020) and later AraELECTRA (Antoun, Baly, & Hajj, 2021), which are the Arabic versions of BERT and ELECTRA (Clark, Luong, Le and Manning, 2020), respectively. The two datasets by Mozannar 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 (Atef, Mattar, Sherif, Elrefai, & Torki, 2020). It is composed of about 17k QA pairs for 3381 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 (Seo, Kembhavi, Farhadi, & Hajishirzi, 2016) as MRC readers. The last datasets to overview are two multilingual MRC datasets, each having a fair share of Arabic questions. The TyDi QA (Clark et al., 2020) and MLQA (Lewis, Oguz, Rinott, Riedel and Schwenk, 2020) 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 (Azmi & Alshenaifi, 2017) and the DAWQAS (Ismail & Homsi, 2018) 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.

2.3. Machine reading comprehension

MRC has been recently fueled by the success of transformer-based (Vaswani et al., 2017) pre-trained language models, exemplified by the phenomenal success of BERT (Devlin et al., 2019) and BERT-like models (Clark, Luong et al., 2020; Liu et al., 2020) 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 Peters et al., 2018; Radford, Narasimhan, Salimans, & Sutskever, 2018), 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 (Yang, Dai et al., 2019), RoBERTa (Liu et al., 2020), GPT-3 (Brown et al., 2020), ELECTRA (Clark, Luong et al., 2020), BART (Lewis et al., 2020), SpanBERT (Joshi et al., 2020) and DeBERTa (He, Liu, Gao, & Chen, 2021) 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 (Joshi et al., 2020).

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 (Khashabi, Chaturvedi, Roth, Upadhyay, & Roth, 2018), DROP (Dua et al., 2019), QUOREF (Dasigi, Liu, Marasović, Smith, & Gardner, 2019), and WikiHowQA (Cui, Hu, & Hu, 2021). 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 (Yu et al., 2018), 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 et al. (2019) and Hu, Peng, Huang, and Li (2019) 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, Globerson, and Berant (2020) proposed an approach that casts the extractive multi-span prediction problem as a sequence tagging task, in 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 et al. (2021) 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 (Dua et al., 2019; Hu et al., 2019; Segal et al., 2020) 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 2.2) are more adequate for single-span extractive MRC.

3. Developing *QRCD*

The Qur'anic Reading Comprehension Dataset, denoted as *QRCD*, is an extension of the AyaTEC dataset (Malhas & Elsayed, 2020) for the task of Machine Reading Comprehension (MRC) on the holy Qur'an. As such, we introduce a short description of AyaTEC before introducing the procedure adopted for developing *QRCD*.

3.1. Original AyaTEC dataset

AyaTEC is a fully re-usable verse-based test collection for Arabic question answering on the Holy Qur'an. It was developed by Malhas and Elsayed (2020) to provide a common experimental testbed for evaluating and fairly comparing the performance of Arabic question answering systems on the Holy Qur'an. In its current version, AyaTEC-v1.1[9] is composed of 1747 QA pairs for 207

[7] There are MRC approaches that require multi-sentence reasoning to answer *single-answer* questions, such as Richardson, Burges, and Renshaw (2013) and Yang, Zhang, and Zhao (2020).

[8] Cui et al. (2021) also applied ListReader on their introduced Chineze WebQA dataset.

[9] http://qufaculty.qu.edu.qa/telsayed/datasets/.

questions. The answers in AyaTEC are verse-based, where each answer is composed of one or more consecutive Qur'anic verses. AyaTEC is fully reusable in the sense that *all* Qur'anic verses that directly answer a question were extracted by Qur'an specialists.

All questions in AyaTEC are expressed in Modern Standard Arabic (MSA) covering a diverse set of 11 Qur'an topics. The questions, collected from different sources, comprise factoid and non-factoid questions that were categorized into three abstract types: *single-answer* questions, *multi-answer* questions, and *no-answer* questions. Three Qur'an specialists annotated the initially-extracted potential answers of all the questions in AyaTEC as either *direct*, *indirect* or *incorrect*. As defined by Malhas and Elsayed (2020), a *direct* answer responds to a question *explicitly* and its context is *consistent* with that of the question. In contrast, an *indirect* answer can either be an answer responding to the question *explicitly* but its context is *inconsistent* with the context of the question, or an answer responding to the question *implicitly* with its context being *consistent* with that of the question. Finally, an answer is *incorrect* if it does not answer the corresponding question.

3.2. Extending AyaTEC

In this section, we describe the procedure for developing *QRCD*. *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. 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 Fig. 1-(a) and (c). A *multi-answer* question is the one whose answers are composed of several components (such as *why* questions) in two or more different answer spans (in distant or contiguous verses) in the accompanying Qur'anic passage, as shown in Fig. 1-(b).

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.

Overall, *QRCD* is composed of 1093 question-passage pairs; 939 of which are single-answer questions and the remaining 154 are multi-answer questions (Table 1). With 14% of the questions in *QRCD* being multi-answer questions, this poses an additional challenge to the reading comprehension task.

3.2.1. 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,[10] which is a printed edition that clusters the verses of each chapter into topics. We recruited two annotators through UpWork[11] to extract the start and end verse numbers to which each topic cluster 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[12] 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 (2020) 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.

3.2.2. 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 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 3.3, we further discuss the inter-annotator agreement and mismatches among the annotators.

The final number of answer spans extracted for the 1093 questions (or question-passage pairs) was 1337 with an average size of eight words per span. Their distribution across question types are shown in Table 1.

[10] http://archive.org/details/Quran_Tafseel-Mawdo.

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

[12] https://tanzil.net/download/.

Table 1
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

3.3. 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 (Sim & Wright, 2005). 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 (Malhas & Elsayed, 2020), 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 Landis and Koch (1977), 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 (Malhas & Elsayed, 2020).

4. Developing CL-AraBERT

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 (Devlin et al., 2019). This is mainly attributed to the unsupervised (or self-supervised) nature of LM pre-training, the ubiquitous presence of unlabeled text to train on, and the advent of transformer-based models such as GPT (Radford et al., 2018) and BERT 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 (Antoun et al., 2020). 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 et al. (2019), 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 and AraBERT models. In Section 4.1, 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 4.2 and 4.3, respectively.

4.1. Classical Arabic data for pre-training

Devlin et al. (2019) 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 et al. (2020) 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 (2016), and (ii) the OSIAN corpus by Zeroual, Goldhahn, Eckart, and Lakhouaja (2019). As such, all their pre-training data resources were in MSA. The size of their final pre-training dataset was ~24 GB 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 (Abdelali, Darwish, Durrani, & Mubarak, 2016) 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 (Antoun et al., 2020). We have chosen to use AraBERTv0.1.

As AraBERT was pre-trained using MSA resources only, we used the OpenITI corpus (Romanov & Seydi, 2019) 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,[13] 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[14] and Al-Jami' Al-Kabir,[15] 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.[16] 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.

4.2. 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 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[17] used in BERT to learn the vocabulary and generate the WordPiece embeddings (Wu et al., 2016). 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,[18] 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. Devlin et al. (2019) defined a "sentence" as any span of consecutive text (rather than a usual linguistic sentence), and defined a "sequence" 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 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 (Antoun et al., 2020; Devlin et al., 2019).

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 (Antoun et al., 2020; Devlin et al., 2019).

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 ~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 et al.). 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.[19] 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.

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

[14] https://shamela.ws/.

[15] 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.

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

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

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

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

4.3. 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 1), we leverage cross-lingual transfer learning by using the Arabic SQuAD and ARCD question answering datasets by Mozannar et al. (2019) 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 1395 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 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 (Devlin et al., 2019). 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 Section 6.1.

5. Evaluation measures

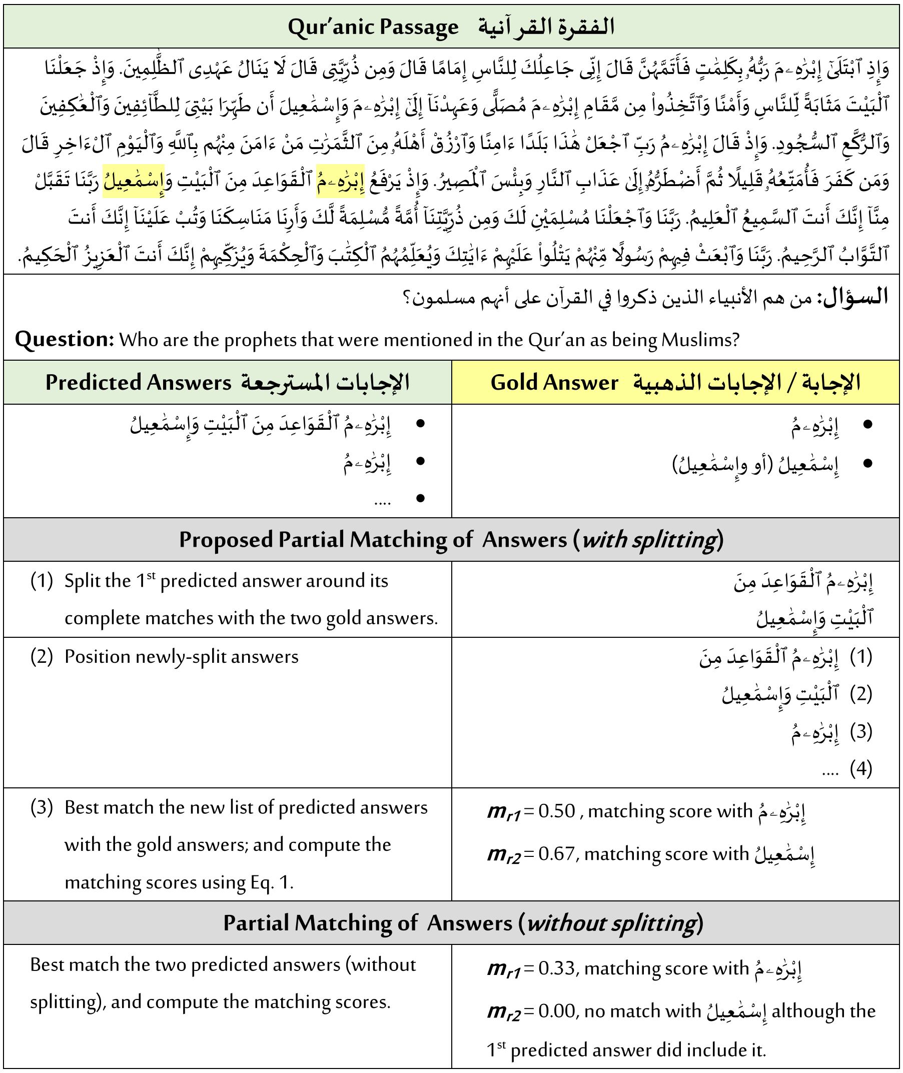
Performance evaluation of an extractive MRC system over a question related to a given Qur'anic passage should not be confined to one predicted answer only, especially for multi-answer questions. Therefore, we expect the *ideal* MRC system to return *all* correct answers *exclusively* (i.e., only the 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).

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*) (Chen et al., 2017; Rajpurkar et al., 2016; Zeng, Li, Li, Hu, & Hu, 2020). 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 et al. (2019) 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 et al. (2018) 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. Fig. 3 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 5.1.

To address the above issues and be able to use a ranked-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 5.1); and adapt the traditional Average Precision (*AP*) rank-based measure (Kishida, 2005) 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 5.2 and 5.3, respectively).[20] The traditional *EM* and token-level F_1 evaluation measures are also adopted, but for single-answer questions only.

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



<table>
<tr><th colspan="2">Qur'anic Passage الفقرة القرآنية</th></tr>
<tr><td colspan="2">وَإِذِ ٱبْتَلَىٰٓ إِبْرَٰهِۦمَ 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 إِبْرَٰهِۦمَ مُصَلًّى وَعَهِدْنَآ إِلَىٰٓ إِبْرَٰهِۦمَ وَإِسْمَٰعِيلَ أَن طَهِّرَا بَيْتِىَ لِلطَّآئِفِينَ وَٱلْعَٰكِفِينَ
وَٱلرُّكَّعِ ٱلسُّجُودِ. وَإِذْ قَالَ إِبْرَٰهِۦمُ 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 إِبْرَٰهِۦمُ 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>• إِبْرَٰهِۦمُ ٱلْقَوَاعِدَ مِنَ ٱلْبَيْتِ وَإِسْمَٰعِيلُ
• إِبْرَٰهِۦمُ
• ....</td><td>• إِبْرَٰهِۦمُ
• 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>إِبْرَٰهِۦمُ ٱلْقَوَاعِدَ مِنَ
ٱلْبَيْتِ وَإِسْمَٰعِيلُ</td></tr>
<tr><td>(2) Position newly-split answers</td><td>(1) إِبْرَٰهِۦمُ ٱلْقَوَاعِدَ مِنَ
(2) ٱلْبَيْتِ وَإِسْمَٰعِيلُ
(3) إِبْرَٰهِۦمُ
(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 إِبْرَٰهِۦمُ
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 إِبْرَٰهِۦمُ
m_{r2} = 0.00, no match with إِسْمَٰعِيلُ although the 1st predicted answer did include it.</td></tr>
</table>

Fig. 3. 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.

We note that rank-based measures were used sparingly for evaluation (Baradaran et al., 2020) over single-answer questions, but they were mainly applied in sentence or answer selection (rather than span extraction) tasks and without integrating partial matching (Min, Zhong, Socher, & Xiong, 2018; Wang, Guo, Liu, He, & Zhao, 2016).

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.

5.1. 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. 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 procedure is presented in Fig. 3.

We note that partial matching *with splitting* induces a ripple effect on the rank order of subsequent predicted answers (as shown in Fig. 3), 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 Fig. 3). 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 (2020) 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.

$$m_r = \max_{a \in A} F_1(r, a) \tag{1}$$

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. Fig. 3 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.

5.2. 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 et al. (2016) to the top predicted answer against its ground truth answer.

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

$$F_1@1(R) = m_{r_1} \tag{2}$$

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}.

$$EM(R) = \begin{cases} 1 & \text{if } m_{r_1} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

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

5.3. 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.[21] *pAP* is defined as follows:

$$pAP(R) = \frac{1}{|A|} \sum_{K=1}^{|R|} \mathbb{1}\{m_{r_K} > 0\} \cdot pPrec@K(R) \tag{4}$$

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 also integrates the concept of partial matching, defined by Malhas and Elsayed (2020) as follows:

$$pPrec@K(R) = \frac{1}{K} \sum_{i=1}^{K} m_{r_i} \tag{5}$$

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 Eq. (1).

[21] Similar to the traditional Average Precision (*AP*) (Kishida, 2005), *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 ranks for which precision is zero).

Table 2
Distribution of question-passage-answer triplets across *QRCD* splits into training and test/holdout datasets.

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

To elaborate more on how the pAP measure is computed and showcase its fairness, Fig. 10 in Appendix 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 5.1.

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

6. Experimental evaluation

In this section, we describe the setup of our experiments, then present the evaluation results and discuss them and their implications in the context of addressing the three research questions listed below (Section 6.2). This is followed by a performance analysis of the best performing model (Section 6.2.4), in which we discuss some failure and success examples to draw insight into future directions to address the identified challenges. We conclude this section with several general implications of this research work (Section 6.3).

RQ1: Does further pre-training with Classical Arabic improve the performance over the MSA-only pre-trained model?
RQ2: 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?
RQ3: How does the fine-tuned CL-AraBERT reader perform on multi-answer questions vs. single-answer questions?

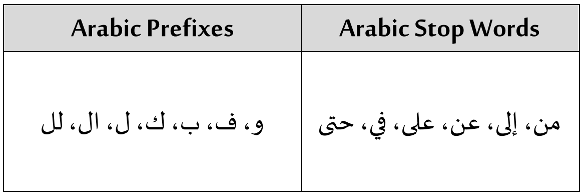
6.1. Experimental setup

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 2. 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. 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.

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 et al. (2016) 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. 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 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 5, 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 note that before applying the partial matching procedure described in Section 5.1 during evaluation, the Farasa tool (Abdelali et al., 2016) 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 beginning of the gold answers during their extraction by the annotators. The prefixes and stopwords that were removed are shown in Fig. 4.



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

Fig. 4. The Arabic prefixes and stopwords removed before comparing the predicted and gold answers during evaluation.

Table 3
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 datasets	Holdout setup	CV setup
		$pAP@10$	$pAP@10$
(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	$\mathbf{51.49}^{abce}$	$\mathbf{53.28}^{abcde}$

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.

6.2. Results and discussion

Tables 3 and 4 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%.

6.2.1. Comparing performance of CL-AraBERT to AraBERT (RQ1)

We start by addressing *RQ1*, which is concerned with observing the effect of further pre-training the MSA pre-trained model with Classical Arabic data. Table 3 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 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 3). 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 (Romanov & Seydi, 2019).

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 CV and the hold-out 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.

Table 4
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. type	Holdout setup			Cross-validation setup		
		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			$\mathbf{47.53}^{abde}$			$\mathbf{47.40}^{abde}$

6.2.2. *Transfer learning from MSA to Classical Arabic (RQ2)*

We address the second research question (RQ2), 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$_{\mathrm{msa}+qrcd}$ 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 3). Similar improvements were also witnessed by CL-AraBERT$_{\mathrm{msa}+qrcd}$ in comparison to CL-AraBERT$_{qrcd}$ as shown in Table 3. 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 (Mozannar et al., 2019).

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$_{\mathrm{msa}+qrcd}$ and CL-AraBERT$_{\mathrm{msa}+qrcd}$ with their counter models that were exclusively fine-tuned using the two MSA datasets, has revealed this gap, especially in the CV setup. AraBERT$_{\mathrm{msa}+qrcd}$ outperformed AraBERT$_{\mathrm{msa}}$ by ~14.9 points on its pAP score (Table 3). Likewise, CL-AraBERT$_{\mathrm{msa}+qrcd}$ outperformed CL-AraBERT$_{\mathrm{msa}}$ by ~13.8 points on its pAP score (Table 3). The pAP scores in the holdout setup have also revealed this difference in performance, but with a lesser extent.

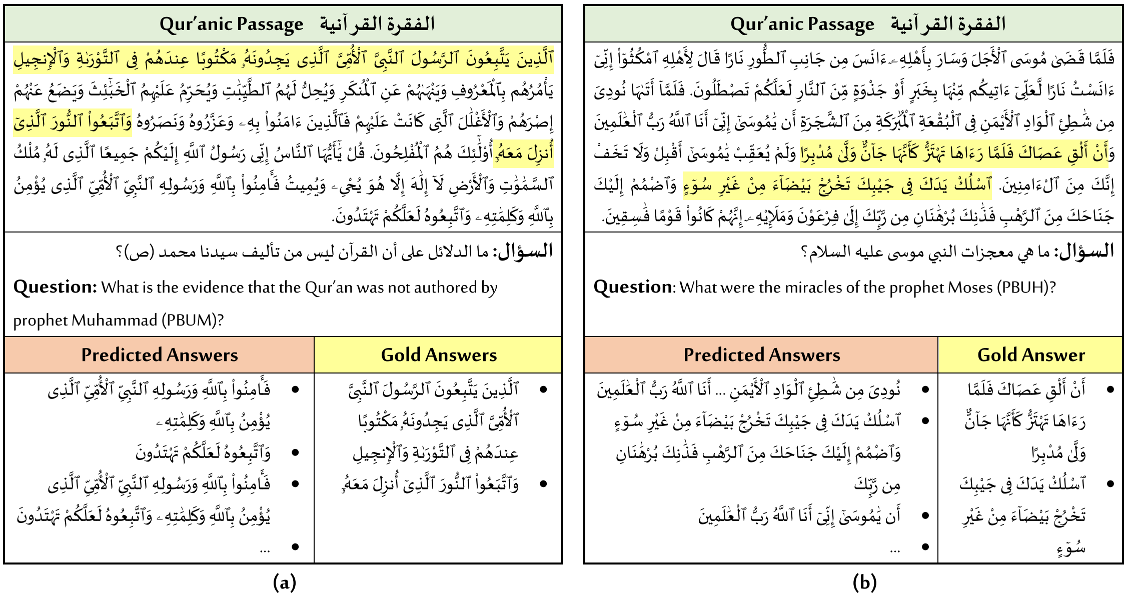
While the performance using MSA-only datasets is fair, the above findings demonstrate the impact of the *QRCD* dataset (as a Classical Arabic resource) in boosting performance of classical and non-classical models, despite its relatively modest size of 1337 question-passage-answer triplets. They also 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 (Sharaf & Atwell, 2012), despite the contemporary western words that found their way into MSA through translation or transliteration.

6.2.3. *Performance across question types (RQ3)*

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

Table 4 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-answer questions. Moreover, multi-answer questions are naturally more challenging, hence typically harder. Again, CL-AraBERT$_{\mathrm{msa}+qrcd}$ 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 4, 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



Qur’anic Passage الفقرة القرآنية

ٱلَّذِينَ يَتَّبِعُونَ ٱلرَّسُولَ ٱلنَّبِىَّ ٱلْأُمِّىَّ 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 ٱلنَّبِىَّ ٱلْأُمِّىَّ 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)

Fig. 5. 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$_{msa+qrcd}$.

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.

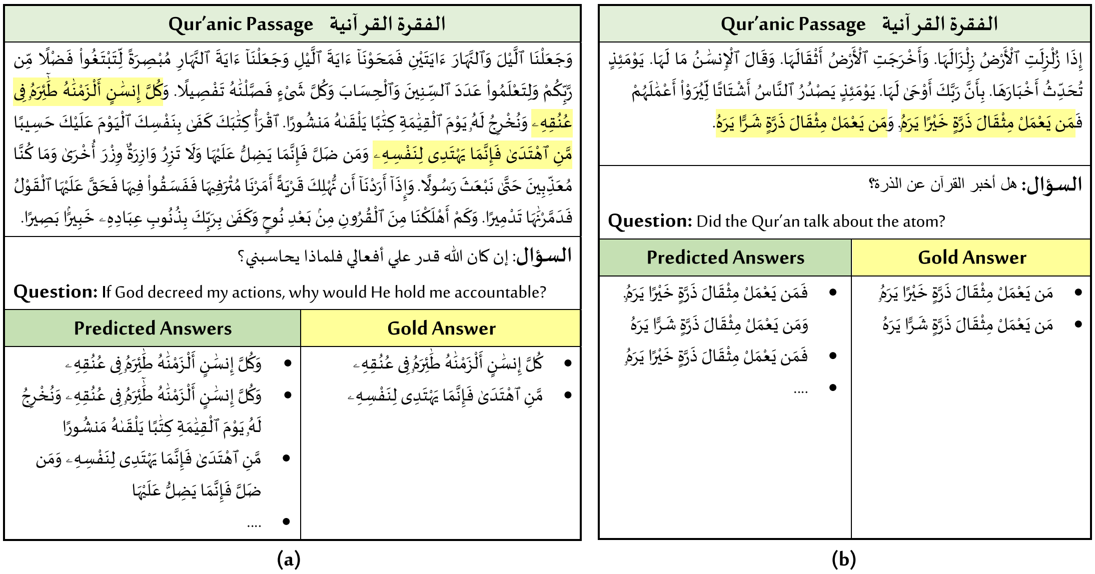
6.2.4. Performance analysis

In this section, we discuss and present several failure and success examples (in Figs. 5 through 9) in an attempt to understand the weaknesses and strengths of the fine-tuned CL-AraBERT$_{msa+qrcd}$ 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. Failure to answer some questions could be attributed to one or more of the following challenges, though CL-AraBERT$_{msa+qrcd}$ 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 to non-factoid questions are mostly evidence-based in the Holy Qur’an. For the multi-answer question in Fig. 5(a), the reader failed to return the two answers which provide evidence that prophet Muhammad (PBUH) did not author the Qur’an, while in Fig. 6(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 Figs. 7(b) and 6(b), respectively.
We note that some of the examples mentioned above (such as Figs. 7(b) and 6(a)) may also demonstrate one or more of the challenges described in 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 Fig. 7(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 Fig. 8(a) and (c), despite the relatively large distance between the antecedents (highlighted in yellow) and the reference expressions (highlighted in blue) in the respective Qur’anic passages; the distance reached 2 verses with ~78 words for the anaphoric (i.e., coreference) expression in Fig. 8(a), and 2 verses with ~33 words for the expression in Fig. 8(c).



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.

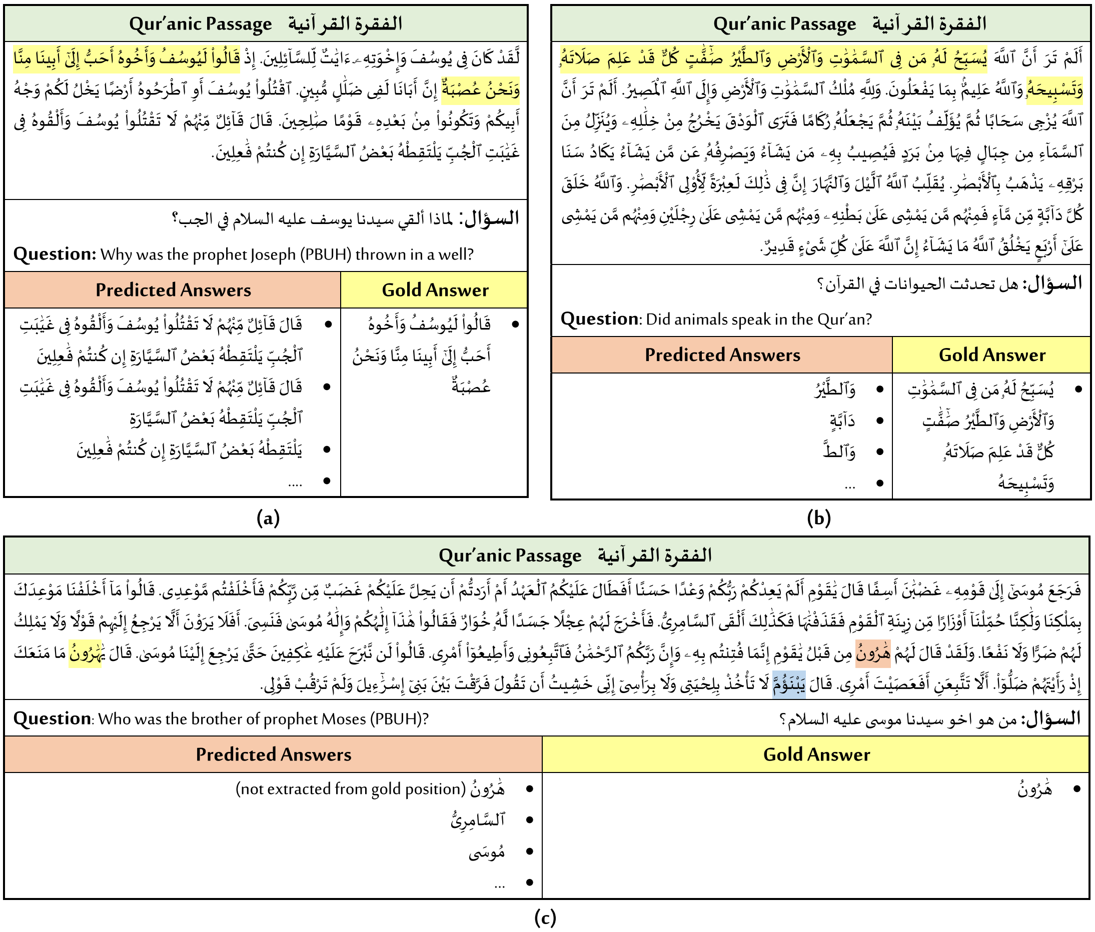
السؤال: هل أخبر القرآن عن الذرة؟

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)

Fig. 6. Two **success** examples of **multi-answer** questions correctly answered by CL-AraBERT$_{msa+qrcd}$.



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

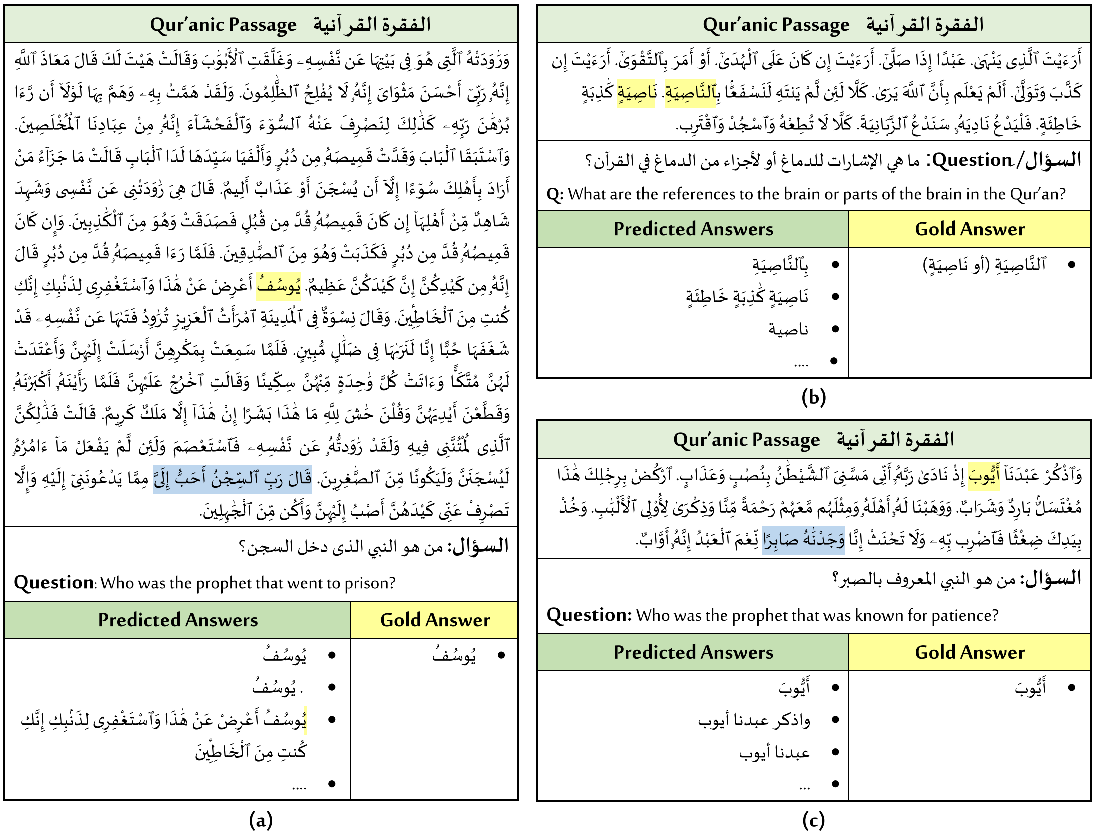
فَرَجَعَ مُوسَىٰٓ إِلَىٰ قَوْمِهِۦ غَضْبَٰنَ أَسِفًا قَالَ يَٰقَوْمِ أَلَمْ يَعِدْكُمْ رَبُّكُمْ وَعْدًا حَسَنًا أَفَطَالَ عَلَيْكُمُ ٱلْعَهْدُ أَمْ أَرَدتُّمْ أَن يَحِلَّ عَلَيْكُمْ غَضَبٌ مِّن رَّبِّكُمْ فَأَخْلَفْتُم مَّوْعِدِى. قَالُوا۟ مَآ أَخْلَفْنَا مَوْعِدَكَ بِمَلْكِنَا وَلَٰكِنَّا حُمِّلْنَآ 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)

Fig. 7. Three **failure** examples of **single-answer** questions that were not correctly answered by CL-AraBERT$_{msa+qrcd}$. Text highlighted in blue is the reference expression to the preceding antecedent highlighted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



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)

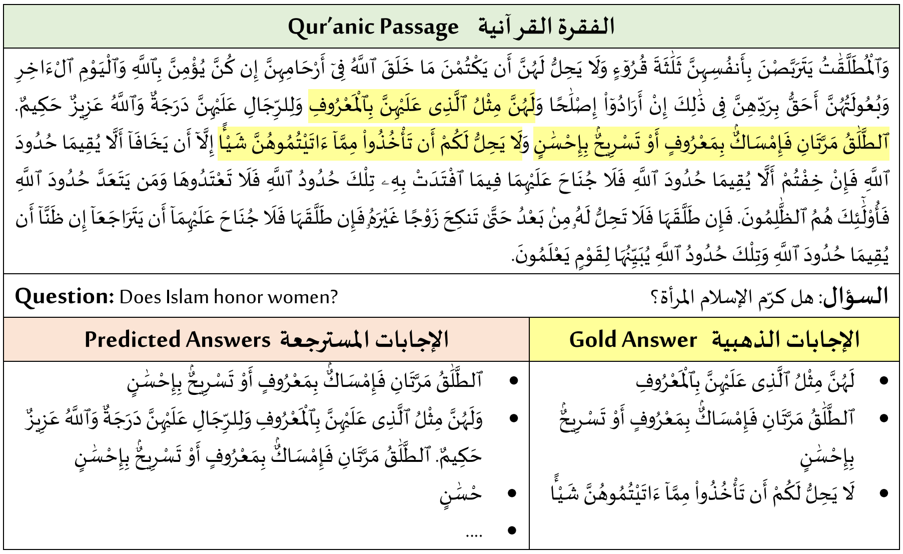
Qur’anic Passage الفقرة القرآنية	
أَرَءَيْتَ ٱلَّذِى يَنْهَىٰ. عَبْدًا إِذَا صَلَّىٰٓ. أَرَءَيْتَ إِن كَانَ عَلَى ٱلْهُدَىٰٓ. أَوْ أَمَرَ بِٱلتَّقْوَىٰٓ. أَرَءَيْتَ إِن كَذَّبَ وَتَوَلَّىٰٓ. أَلَمْ يَعْلَم بِأَنَّ ٱللَّهَ يَرَىٰ. كَلَّا لَئِن لَّمْ يَنتَهِ لَنَسْفَعًۢا بِٱلنَّاصِيَةِ. نَاصِيَةٍ كَٰذِبَةٍ 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)

Fig. 8. Three **success** examples of **single-answer** questions correctly answered by CL-AraBERT$_{msa+qrcd}$. Text highlighted in blue represent reference expressions to the respective preceding antecedents highlighted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Qur’anic Passage الفقرة القرآنية	
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.	
Question: Does Islam honor women? **السؤال**: هل كرّم الإسلام المرأة؟	
Predicted Answers الإجابات المسترجعة	**Gold Answer الإجابات الذهبية**
• 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 •	• لَهُنَّ مِثْلُ ٱلَّذِى عَلَيْهِنَّ بِٱلْمَعْرُوفِ • ٱلطَّلَٰقُ مَرَّتَانِ فَإِمْسَاكٌۢ بِمَعْرُوفٍ أَوْ تَسْرِيحٌۢ بِإِحْسَٰنٍ • لَا يَحِلُّ لَكُمْ أَن تَأْخُذُوا۟ مِمَّآ ءَاتَيْتُمُوهُنَّ شَيْـًٔا

Fig. 9. An example that demonstrates two types of partial failure: missing to predict/extract the third gold answer component, and predicting an answer span that includes a non-essential word/phrase.

(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 Fig. 5(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 Fig. 7(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, Fig. 8(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, Fig. 6(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 (Malhas & Elsayed, 2020). 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 5.1. For the factoid and single-answer question in Fig. 7(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 Fig. 9); (ii) partially predicting an answer, while leaving out an essential word/phrase (e.g., the first predicted answer in Fig. 7(b)); or (iii) predicting an answer span that includes a non-essential word/phrase (e.g., the second predicted answers in Figs. 9 and 5(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 (2012), which is a large corpus of the Qur'an annotated with pronominal anaphora.

6.3. General implications

Our work has several theoretical and practical general implications summarized below.

- **QRCD encouraging further research on the problem**. We note that the attained scores by the best performing CL-AraBERT$_{\text{msa}+qrcd}$ model are relatively modest, implying that the *QRCD* dataset is challenging enough to hopefully trigger further development of state-of-the-art 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.
- **Major step towards retriever-reader QA models on Holy Qur'an**. We also perceive CL-AraBERT$_{\text{msa}+qrcd}$ (or any future fine-tuned classical reader model) as an integral component towards the development of a closed-domain retriever-reader QA model on the Holy Qur'an, where the model gets an MSA question only and aims to find the answer anywhere in the Holy Qur'an.
- **Leveraging CL-AraBERT for other 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 5.1, can also be applied to other rank-based measures, such as $nDCG$ and *Reciprocal Rank*.

7. Conclusion and future work

Motivated by the success of transformer-based neural models on reading comprehension tasks, and the lack of Arabic datasets and systems for extractive MRC on the Holy Qur'an, we introduced *QRCD* as the first Qur'anic Reading Comprehension Dataset. It is composed of 1337 question-passage-answer triplets for 1093 questions that are coupled with their corresponding Qur'anic passages. As the questions in *QRCD* include multi-answer (besides single-answer) questions, the dataset presents an additional challenge to the task. We also introduced CL-AraBERT (CLassical AraBERT), which is a further pre-trained version of AraBERT 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 NLP tasks on CA text such as the Holy Qur'an. Finally, we fined-tuned CL-AraBERT as a reader using two MRC datasets in MSA, prior to fine-tuning it using our *QRCD* dataset. Casting the problem as a cross-lingual transfer learning task from MSA to CA was necessary, not only to address the challenge of having the questions posed in MSA and their answers in Qur'anic CA, but also to overcome the modest size of the *QRCD* dataset.

The need to evaluate the CL-AraBERT reader, or any other extractive MRC system, on multi-answer questions was an eye-opener to the absence in the literature of rank-based measures that can fairly integrate partial matching in the evaluation. 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 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. Moreover, it has also outperformed the baseline over both types of questions. Furthermore, 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.

With the CL-AraBERT reader model attaining pAP (our newly adapted performance measure) of 51.49 and 53.28 in the holdout and cross validation experimental setups over *QRCD*, respectively, we believe there is still room for improving its performance. As such, we make the CL-AraBERT model and the *QRCD* training set publicly available to the research community hoping to elicit state-of-the-art research in Arabic MRC and NLP on the Holy Qur'an and Classical Arabic text, such as Hadith, Exegeses of Qur'an and beyond. A future direction worth exploring is the use of diacritized Arabic text in pre-training and fine-tuning the reader models to study their effect in disambiguating the meaning of Qur'anic verses, and enhancing multi-verse reasoning. On the other hand, research is also needed to explore if recent advances in transformer-based pre-trained language models (that leverage context) are at least on par with the use of diacritics in word sense disambiguation. Another interesting future path for our task is to develop MRC systems that abstain from answering rather than providing a wrong answer. Another future direction that could be considered for enhancing performance is to use the more recent released versions of AraBERT (AraBERTv0.2 base and large).[22] Alternatively, other Arabic BERT-like or transformer-based models that were trained on MSA resources, such as ARBERT (Abdul-Mageed, Elmadany, et al., 2021) and AraELECTRA (Antoun et al., 2021), are worth pre-training using the Classical Arabic corpus.

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 et al. (2021) discuss the caveats and potential pitfalls in the Qur'anic NLP research that we should be wary of.

CRediT authorship contribution statement

Rana Malhas: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper, Writing the research questions, Designing the experiments, Supervising the annotation process, Developing and Implementing the system, Conducting the experiments, Evaluating performance. **Tamer Elsayed:** Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper, Writing the research questions, Designing the experiments.

Data availability

We make our pre-trained model CL-AraBERT, the QRCD dataset, and evaluation script publicly available at our repository: https://github.com/RanaMalhas/QRCD.

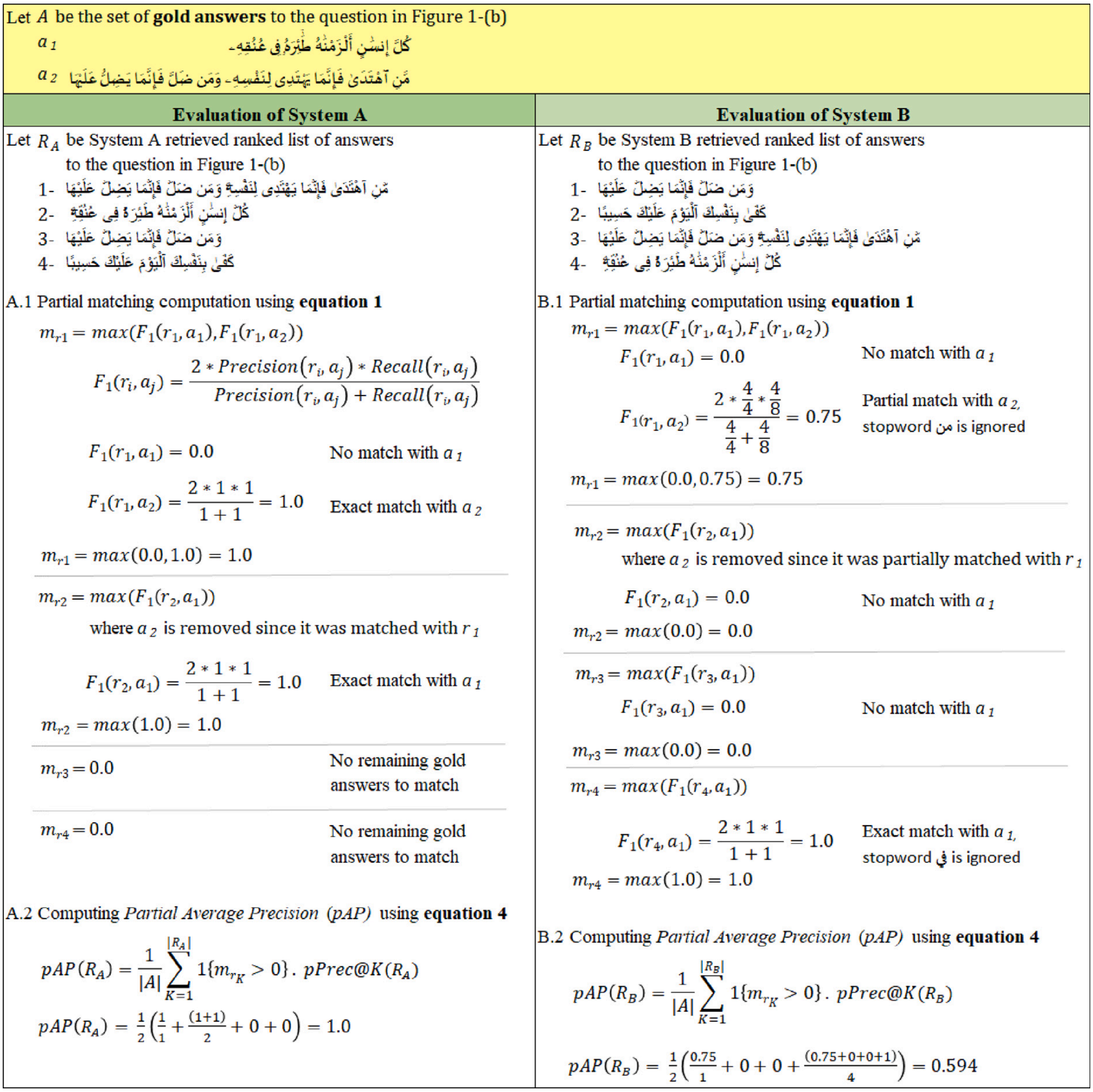
Acknowledgments

We would like to thank TensorFlow Research Cloud (TFRC) for their valuable support in providing us with free access to their cloud TPUs. This research work was partially funded by Qatar University through grant number QUST-2-CENG-2020-20. Open Access funding is provided by the Qatar National Library.

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

Appendix. Example of pAP evaluation

In this appendix, Fig. 10 presents a full example on how the proposed rank-based measure (Partial Average Precision) pAP is computed. The example compares the performance of two different systems given the same 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-(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-(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 using **equation 1**

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

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

$F_1(r_1, a_1) = 0.0$ 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 4**

$$pAP(R_A) = \frac{1}{|A|} \sum_{K=1}^{|R_A|} 1\{m_{r_K} > 0\} . \; pPrec@K(R_A)$$

$$pAP(R_A) = \frac{1}{2}\left(\frac{1}{1} + \frac{(1+1)}{2} + 0 + 0\right) = 1.0$$

Evaluation of System B

Let R_B be System B retrieved ranked list of answers to the question in Figure 1-(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 1**

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

$F_1(r_1, a_1) = 0.0$ 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 4**

$$pAP(R_B) = \frac{1}{|A|} \sum_{K=1}^{|R_B|} 1\{m_{r_K} > 0\} . \; pPrec@K(R_B)$$

$$pAP(R_B) = \frac{1}{2}\left(\frac{0.75}{1} + 0 + 0 + \frac{(0.75+0+0+1)}{4}\right) = 0.594$$

Fig. 10. Full example of how pAP evaluation measure is computed given the returned answers of two different systems on the same question.

References

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). 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* (pp. 11–16). Association for Computational Linguistics.

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N., et al. (2014). Al-Bayan: an Arabic question answering system for the holy quran. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing* ANLP, (pp. 57–64).

Abdul-Mageed, M., Elmadany, A., et al. (2021). 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)* (pp. 7088–7105).

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2012). IDraaq: New arabic question answering system based on query expansion and passage retrieval. In *Proceedings of CLEF 2012 evaluation labs and workshop - Working notes papers*. CELCT.

Akour, M., Abufardeh, S., Magel, K., & Al-Radaideh, Q. (2011). QArabPro: A rule based question answering system for reading comprehension tests in Arabic. *American Journal of Applied Sciences*, *8*(6), 652–661.

Al-Azami, M. (2020). *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.

Alqahtani, M., & Atwell, E. (2018). Annotated corpus of Arabic Al-Quran question and answer.

Alwaneen, T. H., Azmi, A. M., Aboalsamh, H. A., Cambria, E., & Hussain, A. (2021). Arabic question answering system: a survey. *Artificial Intelligence Review*, 1–47.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for arabic language understanding. In *LREC 2020 workshop language resources and evaluation conference 11–16 May 2020* (p. 9).

Antoun, W., Baly, F., & Hajj, H. (2021). AraELECTRA: Pre-training text discriminators for arabic language understanding. In *Proceedings of the sixth Arabic natural language processing workshop* (pp. 191–195). Association for Computational Linguistics.

Atef, A., Mattar, B., Sherif, S., Elrefai, E., & Torki, M. (2020). AQAD: 17,000+ arabic questions for machine comprehension of text. In *2020 IEEE/ACS 17th international conference on computer systems and applications* AICCSA, (pp. 1–6). IEEE.

Azmi, A. M., & Alshenaifi, N. A. (2017). LEMAZA: An Arabic why-question answering system. *Natural Language Engineering*, *23*(6), 877–903.

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

Bakari, W., Trigui, O., & Neji, M. (2014). Logic-based approach for improving arabic question answering. In *2014 IEEE international conference on computational intelligence and computing research* (pp. 1–6). IEEE.

Baradaran, R., Ghiasi, R., & Amirkhani, H. (2020). A survey on machine reading comprehension systems. *Natural Language Engineering*, 1–50.

Bashir, M. H., Azmi, M. A., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., et al. (2021). Arabic natural language processing for qur'anic research: A systematic review.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, D. (2018). *Neural reading comprehension and beyond* (Ph.D. thesis), Stanford University.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 1870–1879). Association for Computational Linguistics.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., et al. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2174–2184). Association for Computational Linguistics.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., et al. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, *8*, 454–470.

Clark, C., & Gardner, M. (2018). Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 845–855). Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International conference on learning representations* (p. 18).

Cui, P., Hu, D., & Hu, L. (2021). ListReader: Extracting list-form answers for opinion questions. arXiv preprint arXiv:2110.11692.

Dasigi, P., Liu, N. F., Marasović, A., Smith, N. A., & Gardner, M. (2019). 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)* (pp. 5925–5932).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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)* (pp. 4171–4186). Association for Computational Linguistics.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019). 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)* (pp. 2368–2378). Association for Computational Linguistics.

El-Khair, I. A. (2016). 1.5 Billion words Arabic corpus. arXiv preprint arXiv:1611.04033.

Ezzeldin, A. M., Kholief, M. H., & El-Sonbaty, Y. (2013). ALQASIM: Arabic language question answer selection in machines. In *International conference of the cross-language evaluation forum for European languages* (pp. 100–103). Springer.

Hakkoum, A., & Raghay, S. (2016). Semantic Q&A system on the Qur'an. *Arabian Journal for Science and Engineering*, *41*(12), 5205–5214.

Hamdelsayed, M., & Atwell, E. (2016). Islamic applications of automatic question-answering. *Journal of Engineering and Computer Science*, *17*(2), 51–57.

Hamoud, B., & Atwell, E. (2016). 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*, *4*(4), 20–29.

Hamoud, B., & Atwell, E. (2017). Evaluation corpus for restricted-domain question-answering systems for the holy Quran. *International Journal of Science and Research*, *6*(8), 1133–1138.

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International conference on learning representations*.

Hu, M., Peng, Y., Huang, Z., & Li, D. (2019). 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)* (pp. 1596–1606).

Ismail, W. S., & Homsi, M. N. (2018). DAWQAS: A dataset for arabic why question answering system. *Procedia Computer Science*, *142*, 123–131.

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

Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). 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)* (pp. 1601–1611). Association for Computational Linguistics.

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018). 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)* (pp. 252–262).

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

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., et al. (2018). The narrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317–328.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794). Association for Computational Linguistics.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). 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* (pp. 7871–7880). Association for Computational Linguistics.

Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Schwenk, H. (2020). MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7315–7330). Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2020). RoBERTa: A robustly optimized BERT pretraining approach. In *International conference on learning representations*.

Malhas, R., & Elsayed, T. (2020). AyaTEC: Building a reusable verse-based test collection for Arabic question answering on the Holy Qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(6), 1–21.

Malhas, R., Mansour, W., & Elsayed, T. (2022). 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)* (pp. 79–87).

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281.

Min, S., Zhong, V., Socher, R., & Xiong, C. (2018). 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)* (pp. 1725–1735).

Mozannar, H., Maamary, E., El Hajal, K., & Hajj, H. (2019). Neural arabic question answering. In *Proceedings of the fourth Arabic natural language processing workshop* (pp. 108–118). Association for Computational Linguistics.

Newman, D. L. (2013). The Arabic literary language: the nahda and beyond. In *The Oxford handbook of arabic linguistics* (p. 472). Oxford University Press.

Peñas, A., Hovy, E. H., Forner, P., Rodrigo, Á., Sutcliffe, R. F., Forascu, C., et al. (2012). Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation.. In *CLEF (Notebook papers/labs/workshop)* (pp. 1–20).

Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Sutcliffe, R., & Morante, R. (2013). QA4MRE 2011–2013: Overview of question answering for machine reading evaluation. In *International conference of the cross-language evaluation forum for European languages* (pp. 303–320). Springer.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. In *NAACL*. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*: *Technical Report*, OpenAI.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392). Association for Computational Linguistics.

Richardson, M., Burges, C. J., & Renshaw, E. (2013). 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* (pp. 193–203).

Romanov, M., & Seydi, M. (2019). *OpenITI: a Machine-readable corpus of islamicate texts*. Zenodo.

Saad, M. K., & Ashour, W. M. (2010). Osac: Open source arabic corpora. In *6th archeng int. symposiums, EEECS, Vol. 10*.

Segal, E., Efrat, A., Shoham, M., Globerson, A., & Berant, J. (2020). A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing* EMNLP, (pp. 3074–3080).

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

Sharaf, A.-B. M., & Atwell, E. (2012). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC* (pp. 130–137). Citeseer.

Shmeisani, H., Tartir, S., Al-Na'ssaan, A., & Naji, M. (2014). Semantically answering questions from the holy quran. In *International conference on islamic applications in computer science and technology* (pp. 1–8).

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems, Vol. 30* (pp. 5998–6008). Curran Associates, Inc..

Wang, B., Guo, S., Liu, K., He, S., & Zhao, J. (2016). Employing external rich knowledge for machine comprehension. In *IJCAI*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., et al. (2019). 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)* (pp. 72–77). Association for Computational Linguistics.

Yang, J., Zhang, Z., & Zhao, H. (2020). Multi-span style extraction for generative reading comprehension. arXiv preprint arXiv:2009.07382.

Yatskar, M. (2019). A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. 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)* (pp. 2318–2323). Association for Computational Linguistics.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., et al. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541.

Zeng, C., Li, S., Li, Q., Hu, J., & Hu, J. (2020). A survey on machine reading comprehension—Tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, *10*(2121), 7640.

Zeroual, I., Goldhahn, D., Eckart, T., & Lakhouaja, A. (2019). OSIAN: Open source international Arabic news corpus-preparation and integration into the CLARIN-infrastructure. In *Proceedings of the fourth Arabic natural language processing workshop* (pp. 175–182).

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