Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an

Rana Malhas
Qatar University
rana.malhas@qu.edu.qa

Watheq Mansour
The University of Queensland*
w.mansour@uq.edu.au

Tamer Elsayed
Qatar University
telsayed@qu.edu.qa

Abstract

Motivated by the need for intelligent question answering (QA) systems on the Holy Qur'an and the success of the first Qur'an Question Answering shared task (Qur'an QA 2022 at OSACT 2022), we have organized the second version at ArabicNLP 2023. The Qur'an QA 2023 is composed of two sub-tasks: the passage retrieval (PR) task and the machine reading comprehension (MRC) task. The main aim of the shared task is to encourage state-of-the-art research on Arabic PR and MRC on the Holy Qur'an. Our shared task has attracted 9 teams to submit 22 runs for the PR task, and 6 teams to submit 17 runs for the MRC task. In this paper, we present an overview of the task and provide an outline of the approaches employed by the participating teams in both sub-tasks.

1 Introduction

The timeless and sacred Qur'an will never cease to attract the interest and inquisition of millions of Muslims and non-Muslims for its profound teachings, legislation, and fertile knowledge. Such inquisitions may be driven by learning, curiosity, or skepticism. The Qur'an is composed of 114 Surahs and 6,236 verses (Ayas) of different lengths, with a total of about 80k words. The words, revealed more than 1,400 years ago, are in Classical Arabic.

Extractive question answering (QA) approaches are being formulated in the literature as machine reading comprehension (MRC) tasks (Chen, 2018). Given a passage of text, a system is evaluated based on its ability to correctly answer a set of questions over the given text. We believe that the resurgence of the MRC field should be harnessed to address the timeless interest in the Holy Qur'an and the information needs of its inquisitors and knowledge seekers (Bashir et al., 2022). This has motivated the inception of the first Qur'an Question Answering shared task, Qur'an QA 2022 at OSACT 2022 Workshop (Malhas et al., 2022).

Although MRC systems are relieved from the task of passage retrieval (i.e., the task of retrieving candidate passages that potentially contain answers to a given question) to purely focus on inference and reasoning for answer extraction, the retriever component remains an integral contributor to the performance of end-to-end extractive QA systems that adopt a retriever-reader architecture (Zhu et al., 2021). Prevalent search/retrieval systems on the Holy Qur'an are either keyword-based, semantic-based, or a hybrid of both paradigms. Semantic-based approaches are predominantly ontology-based with almost no use of state-of-the-art approaches such as dense retrieval (Karpukhin et al., 2020), generative retrieval (Santos et al., 2020) and beyond (Malhas, 2023) to our knowledge.

To this end, and to build on the success of the first Qur'an QA 2022 shared task (Malhas et al., 2022), we have organized the second Qur'an QA shared task (Qur'an QA 2023) at ArabicNLP 2023. Qur'an QA 2023 comprises a Qur'anic Passage Retrieval (PR) task and a Machine Reading Comprehension (MRC) task. The PR task aims at finding all Qur'anic passages that have potential answers for a given question that is posed in Modern Standard Arabic (MSA). Whereas the MRC task targets the extraction of all answers to a given question from a given qur'anic passage. Each answer must be a *span* of text extracted from the given passage. To make both tasks more challenging, we include questions that have no answers in the Qur'an. Further details about the two tasks are provided in Sections 3 and 4, respectively.

To encourage quality participation in the task, we allotted five awards. The awards for the best and second-best teams in each task are $300 and $200, respectively, provided that their papers are accepted at the conference. The fifth award is $150 allotted for the best paper.

Part of the work on this paper was done while being at Qatar University.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 690–701

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

Figure 1: An example for the PR task: a factoid question with some of its gold (answer-bearing) Qur'anic passages. Answers are highlighted in each passage.

Qur'an QA 2023[1] has attracted 38 and 29 teams to sign up for the PR Task and the MRC Task, respectively. In the final phase, 9 teams participated in the PR task with 22 run submissions, and 6 teams participated in the MRC task with 17 run submissions. Table 1 lists the participating teams per task with their affiliations and team size. Six of them have accepted system description papers as referenced in the table.

The rest of the paper is organized as follows. Section 2 outlines the first version of Qur'an QA. Sections 3 and 4 discuss the PR and MRC tasks, respectively, in detail including the task descriptions, datasets, evaluation setups, results, and analysis of approaches employed by the participating teams. We conclude with final thoughts in Section 5.

2 The Qur'an QA 2022 Shared Task

The Qur'an QA shared task in its first version in 2022[2] (Malhas et al., 2022) only comprised an MRC task that is similar to the MRC task proposed this year, but it was relatively simplified. It was defined as follows: given a Qur'anic passage that consists of consecutive verses in a specific Surah of the Holy Qur'an and a question posed in MSA over that passage, a system is required to extract *any* correct answer *span* to that question (regardless if the question had more than one answer in that passage or only one answer). As such, the main measure used in the performance evaluation of participating systems was partial Reciprocal Rank (pRR) (Malhas and Elsayed, 2020).

Qur'an QA 2022 has attracted 30 teams to sign up for the task. In the final phase, 13 teams participated, with a total of 30 submitted runs on the test set. Ten out of the thirteen teams submitted system description papers, which were peer-reviewed and published in OSACT 2022 (Al-Khalifa et al., 2022).

3 Task A: Passage Retrieval (PR)

In this section, we define the PR task, introduce the dataset, and elaborate on the evaluation setup and teams' results. We conclude this section with an overview of the main methods employed by the participating teams.

3.1 Task Description

The task is defined as follows: Given a free-text question posed in MSA and a collection of passages that cover the Holy Qur'an, the system is required to return a *ranked list* of up to 10 answer-bearing passages (i.e., passages that potentially enclose all the answers to the given question) from this collection. The question can be factoid or non-factoid. An example question is shown in Figure 1.

To make the task more realistic (thus challenging), some questions may not have an answer in the Holy Qur'an. We call them *zero-answer* questions. In such cases, the ideal system should return no

[1] https://sites.google.com/view/quran-qa-2023
[2] https://sites.google.com/view/quran-qa-2022

Team	Tasks	Size	Affiliations
Al-Jawaab (Zekiye and Amroush, 2023)	A, B	2	Koç University, Niuversity
AHJL (Alawwad et al., 2023)	A	4	King Abdulaziz University, Saudi Electronic University, Imam Mohammad Ibn Saud Islamic University (IMSIU), King Saud University
GYM (Mahmoudi and Morshedzadeh, 2023)	A, B	2	Iran University of Science and Technology, University of British Columbia
LKAU23 (Alnefaie et al., 2023)	A, B	5	University of Leeds, King Abdulaziz University
LowResContextQA (Veeramani and Roy, 2023)	B	2	University of California, Los Angeles (UCLA), UoSC
PSUT	A, B	5	Princess Sumaya University for Technology
sabran	A	1	Independent
SSZ	A	3	Qatar University
TCE (Elkomy and Sarhan, 2023)	A, B	2	Tanta University
TERROR	A	1	Helwan University

Table 1: Participating teams in Qur'an QA 2023.

Dataset	%	# Questions	QP Pairs
Training	70%	174	972
Development	10%	25	160
Test	20%	52	427
All	100%	251	1,599

Table 2: Distribution of questions and question-passage (QP) pairs in the PR dataset (AyaTECv1.2)

answers; otherwise, it returns a ranked list of the answer-bearing passages.

3.2 PR Dataset

In this section, we introduce the dataset/test collection used in the PR task. In general, a *test collection* is typically composed of a document collection[3] (the Holy Qur'an passages in our case), a set of queries (questions), and their relevance judgments (Lin and Katz, 2006) (i.e., the gold answers, or the passages that comprise them in our case).

For the PR task, an extended version of the *AyaTEC* dataset/test collection (Malhas and Elsayed, 2020) was used (AyaTECv1.2).[4] It is composed of the Qur'anic Passage Collection (QPC) (Malhas, 2023; Swar, 2007), an augmented set of AyaTEC's original questions (AyaTECv1.1), and their relevance judgments (i.e., the answer-bearing passages for each question).

The QPC was developed by topically segmenting the 114 Qur'anic Surahs of different lengths using the Thematic Holy Qur'an (Swar, 2007),[5] which is a printed edition that clusters the verses of each Surah into topics. This segmentation resulted in a total of 1,266 topical passages.

As for the set of questions, 199 out of the original 207 questions of the AyaTECv1.1 test collection (Malhas and Elsayed, 2020) were used. This set was augmented with 52 new questions for evaluating the systems in the PR task. Overall, we have included a total of 37 zero-answer questions (about 15%) that do not have an answer in the Holy Qur'an. The distribution of the training (70%), development (10%), and test (20%) splits are exhibited in Table 2.

For the additional 52 questions, we adopted the same verse-based answer extraction/annotation methodology used while developing the original AyaTEC dataset. The extraction of potential verse-based answers was conducted by two annotators who are knowledgeable about the Qur'an, while the annotation was conducted by three Qur'an specialists. Further details about the annotation process are provided in (Malhas and Elsayed, 2020). Developing the relevance judgments of the final set of questions over the QPC were generated automatically using the same methodology adopted by Malhas (2023). Each Qur'anic passage in the collection is considered relevant to the question if it happens to comprise any of the gold verse-based answer(s) completely or partially.

[3] In information retrieval, the term "document collection" or "collection" refers to a corpus or dataset (Yates et al., 2021); we use these terms interchangeably.

[4] https://gitlab.com/bigirqu/quran-qa-2023

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

3.3 Evaluation Setup

In this section, we shed light on the setup and methodology followed in evaluating the performance of participating systems.

3.3.1 Leaderboard and Repository

The leaderboard for both the PR and MRC tasks was hosted on CodaLab (Pavao et al., 2023) to allow participants to evaluate their runs and facilitate benchmarking. A participating team is required to submit their results/answers in one file, denoted as a "run file" or a "run" in short. The run should match TREC run format, i.e., having the following columns: ["question-id", "Q0", "passage-id", "rank", "score", "tag"]. Each team is allowed to submit 30 runs on the dev set, but up to 3 runs on the test set. Each run typically constitutes the results of a different system or a model.

To facilitate checking and evaluating runs before their submittal to the leaderboard, we made the submission-checker and evaluation scripts publicly available through the official repository of the shared task.[6] Furthermore, to give participants a reference point over the leaderboard, we opted for BM25 (a simple, yet very common, classical lexical-based retrieval model) as a baseline, and released the code to the same repository.

3.3.2 Evaluation Measures

As the PR task is a classical *ranked retrieval* task, we adopt Mean Average Precision at depth 10 ($MAP@10$) as the main official evaluation measure. We also report the Mean Reciprocal Rank at depth 10 ($MRR@10$) to measure the performance of retrieving *any* answer-bearing passage. The no-answer cases are handled simply by giving full credit to "no answer" system output, and zero otherwise, in both measures.

3.4 Results

Thirty eight teams registered for the PR task. Among these teams, nine participated in the final (test) phase and submitted 22 runs. The teams are officially ranked based on their best performing submitted run. Table 3 demonstrates the performance of all submitted runs in the test phase ranked by MAP@10.

We note that 8 runs from 3 teams outperformed the baseline, whereas the rest were below it. The highest scores of MAP@10 and MRR@10 are *0.2506* and *0.4610*; both were achieved by the *TCE* team (Elkomy and Sarhan, 2023). Figure 4 (in the Appendix) shows the boxplots for all submitted runs on the test queries (questions) to illustrate the performance distribution. The boxplots reveal the diverse performance across the questions for most of the runs.

Team	Run	MAP@10	MRR@10
TCE	M00	**0.2506**	**0.4610**
TCE	A00	0.2464	0.4940
TCE	C00	0.2302	0.4706
AHJL	SG2	0.1995	0.3889
AHJL	SWOP3	0.1318	0.3021
LKAU23	run63	0.1242	0.3750
AHJL	SS1	0.1202	0.2907
LKAU23	run61	0.1166	0.3632
Baseline	*BM25*	0.0904	0.2260
SSZ	run02	0.0804	0.2177
TERROR	new01	0.0789	0.1608
SSZ	un01	0.0784	0.2206
TERROR	new03	0.0739	0.1566
LKAU23	run62	0.0701	0.2047
Al-Jawaab	test	0.0643	0.1609
GYM	GRun1	0.0545	0.1581
TERROR	new02	0.0327	0.0737
GYM	Run0	0.0315	0.1023
PSUT	run3	0.0214	0.0752
GYM	Run2	0.0116	0.0356
PSUT	run2	0.0114	0.0523
sabran	vers01	0.0000	0.0000
Al-Jawaab	trem02	0.0000	0.0000

Table 3: PR evaluation results of all submitted runs ranked by MAP. The team name is removed from the run name to save space. The underlined rows are the median runs.

3.5 Methods and Analysis

In this section, we give an overview of the main approaches adopted by the 9 participating teams in their submitted runs on the test set. We do that in the context of highlighting some of our perceptions and general trends that characterize the participating systems and their submitted runs.

As expected, all systems utilized pre-trained transformer-based Language Models (LMs), two of which used generative (decoder-only) LMs (e.g., GPT), while the remaining systems employed encoder-only LMs (e.g., BERT). The majority of the semantic search/retrieval systems used bi-encoder and cross-encoder architectures either in-

[6] https://gitlab.com/bigirqu/quran-qa-2023

الفقرة القرآنية (74:32-48) Qur'anic Passage
كَلَّا وَٱلْقَمَرِ. وَٱلَّيْلِ إِذْ أَدْبَرَ. وَٱلصُّبْحِ إِذَآ أَسْفَرَ. إِنَّهَا لَإِحْدَى ٱلْكُبَرِ. نَذِيرًا لِّلْبَشَرِ. لِمَن شَآءَ مِنكُمْ أَن يَتَقَدَّمَ أَوْ يَتَأَخَّرَ. كُلُّ نَفْسٍۭ بِمَا كَسَبَتْ رَهِينَةٌ. إِلَّآ أَصْحَٰبَ ٱلْيَمِينِ. فِى جَنَّٰتٍ يَتَسَآءَلُونَ. عَنِ ٱلْمُجْرِمِينَ. مَا سَلَكَكُمْ فِى سَقَرَ. قَالُوا۟ لَمْ نَكُ مِنَ ٱلْمُصَلِّينَ. وَلَمْ نَكُ نُطْعِمُ ٱلْمِسْكِينَ. وَكُنَّا نَخُوضُ مَعَ ٱلْخَآئِضِينَ. وَكُنَّا نُكَذِّبُ بِيَوْمِ ٱلدِّينِ. حَتَّىٰٓ أَتَىٰنَا ٱلْيَقِينُ. فَمَا تَنفَعُهُمْ شَفَٰعَةُ ٱلشَّٰفِعِينَ.
السؤال/ Question: ما هي الدلائل التي تشير بأن الانسان مخير؟
الإجابات الذهبية / Gold Answers:
1. لِمَن شَآءَ مِنكُمْ أَن يَتَقَدَّمَ أَوْ يَتَأَخَّرَ
2. كُلُّ نَفْسٍۭ بِمَا كَسَبَتْ رَهِينَةٌ

Figure 2: An example of the MRC task: a non-factoid question with the answers highlighted in the given passage.

dependently or jointly. Also, ensemble and self-ensemble approaches were employed by many systems to stabilize prediction fluctuations across runs and/or to enhance prediction accuracy through the wisdom of the crowd. For zero-answer questions, the majority of the systems did not explicitly address this challenge.

The three run submissions of the TCE team (Elkomy and Sarhan, 2023) outperformed all other submissions to the PR task (Table 3). TCE's three systems leveraged transfer learning and ensemble learning while training their dual-encoders (bi-encoders) and cross-encoders for ad hoc search (Yates et al., 2021). Their top performing system (with a MAP score of 0.2505) employed an ensemble of CAMel_BERT-CA (Inoue et al., 2021) and AraBERTv0.2-base (Antoun et al., 2020) dual encoder. Each of these encoders was self-ensembled and fine-tuned using three datasets; namely, the Arabic part of the multilingual TyDiQA dataset (Ar_TyDiQA) (Clark et al., 2020), followed by a Tafseer dataset [7] and finally the Task A dataset (AyaTECv1.2). Their second and third best systems employed the self-ensembled AraBERTv0.2-base and CAMel_BERT-CA encoders, respectively. For zero-answer questions, TCE adopted a thresholding mechanism to identify questions with a low cumulative likelihood of having answers in the Holy Qur'an. However, the threshold value should have been tuned rather than being set to approximately equal the percentage of zero-answer questions in AyaTECv1.2 training and development datasets.

The second-best ranked team (AHJL) (Alawwad et al., 2023) employed two semantic search models that were equipped with a translation module to translate a given question to English prior to performing the search. As such, an English translation of the meanings of the Qur'an was used. Given a translated question, the retriever module employs a bi-encoder to retrieve relevant passage candidates, then a cross-encoder is employed as a re-ranker. A zero-shot training setting was adopted. The OpenAI embeddings-based (OpenAI, 2023c) semantic search model (with translation) was their best-performing system (attaining a MAP score of 0.1995) while being able to successfully identify more than half of the zero-answer questions in the test set. OpenAI's best embeddings 'text-embedding-ada-002' model (OpenAI, 2023b) was employed as the bi-encoder (for primary search) and OpenAI's 'text-davinci-003' model as the cross-encoder (for re-ranking). Their second best performing system adopted the SBERT API (Reimers, 2023) that adopts the Sentence-BERT architecture (Reimers and Gurevych, 2019) with translation as well. It employed the 'msmarco-distilbert-base-tas-b' (Reimers, 2023) sentence transformer model as the bi-encoder and 'ms-marco-MiniLM-L-6-v2' (Reimers, 2023) as the cross-encoder. We note that Al-Jawaab team also employed a bi-encoder architecture using the same OpenAI's embeddings used by the AHJL team for their bi-encoder, but their MAP score attained a below-median score of 0.0643.

The third ranked team (LKAU23) (Alnefaie et al., 2023) also adopted the Sentence-BERT architecture for their four Arabic pre-trained LMs (bi-encoders) fine-tuned using AyaTECv1.2 and QRCDv1.2 datasets, respectively. Their best-performing model was CL-AraBERT (Malhas and Elsayed, 2022) which attained a better MAP score (0.1242) than that of ArabicBERT (Safaya et al., 2020), CAMeL-BERT (Inoue et al., 2021), and AraBERT (Antoun et al., 2020). Their second performing model was an ensemble of ArabicBERT

[7] Interpretation resources (Tafseer) from Al-Muyassar and Al-Jalalayn were obtained from Tanzil https://tanzil.net/docs/resources.

and CL-AraBERT that attained a MAP score of 0.1166, which is still better than the median and baseline scores (Table 3).

Among the remaining submitted runs that attained near-median MAP scores (but below the baseline score) belonged to the SSZ and the TERROR teams. The GYM team (Mahmoudi and Morshedzadeh, 2023) attained a below-median MAP score of 0.0545 despite their deployment of an interesting approach that leveraged *unsupervised* fine-tuning of sentence bi-encoders using Transformer-based Sequential Denoising Auto Encoders (TSDAE) (Wang et al., 2021) and Simple Contrastive Learning of Sentence Embeddings (SimCSE) (Gao et al., 2021). The bi-encoder is then fine-tuned using a multi-task learning approach. Their best-performing run employed an AraBERT bi-encoder fine-tuned using the QPC dataset with the TSDAE unsupervised method. Then, it was fine-tuned using Mr. TyDi's Arabic dataset (Zhang et al., 2021) and the Qur'an-passage pairs of AyaTECv1.2 with a multi-task learning approach.

4 Task B: Machine Reading Comprehension (MRC)

In this section, we define the MRC task, present the dataset, and detail the evaluation methodology and results. We conclude with an overview of the main methods employed by the participating teams.

4.1 Task Description

The task is defined as follows: Given a Qur'anic passage that consists of consecutive verses in a specific Surah of the Holy Qur'an, and a free-text question posed in MSA over that passage, a system is required to extract *all* answers to that question that are stated in the given passage (rather than *any* answer as in Qur'an QA 2022). Each answer must be a *span* of text extracted from the given passage. If a question has only one answer in the given passage, it is considered a *single-answer* question, whereas if the question's answer is composed of more than one component/span in the accompanying passage, then the question is considered a *multi-answer* question. We note that the question can be a factoid or non-factoid question. An example is shown in Figure 2.

To make the task more realistic (thus challenging), some questions may not have an answer in the given passage. In such cases, the ideal system should return no answers; otherwise, it returns a ranked list of up to 10 answer spans.

Question Type	QP Pairs				QPA Triplets
	Train	Dev	Test	All	
Multi-answer	134 (14%)	29 (18%)	62 (15%)	224 (14%)	552 (29%)
Single-Answer	806 (81%)	124 (76%)	331 (81%)	1,261 (81%)	1,261 (67%)
Zero-Answer	52 (5%)	10 (6%)	14 (4%)	76 (5%)	76 (4%)
All	992	163	407	1,562	1,889

Table 4: Distribution of question-passage (QP) pairs and question-passage-answer (QPA) triplets by question type in the dataset of Task B (QRCDv1.2)

4.2 MRC Dataset

For the MRC task, an extended version of the Qur'anic Reading Comprehension Dataset (QRCD) (Malhas and Elsayed, 2022) was used (QRCDv1.2). It is composed of the original 1,093 question-passage (QP) pairs in QRCDv1.1, and an augmented set of 62 QP pairs whose questions have no answer in the accompanying passages (nor in the Holy Qur'an). These additional zero-answer questions were paired with hard negative passages retrieved using a BM25 retrieval model. We chose not to pair hard negative passages with the original (single-answer and multi-answer) questions so as not to contaminate the QRCD dataset with non-answer-bearing passages to questions which the Holy Qur'an does have an answer for.

To evaluate the systems in the MRC task, 407 additional QP pairs were included in QRCD, whose questions are the same new questions introduced to the dataset of the PR task (AyaTECv1.2) in Section 3.2 above. Fourteen (14) out of the 407 QP pairs have no answer in the Holy Qur'an; thus, they were also paired with hard negatives. The distribution of the training, development, and test sets are shown in Table 4.

For the additional QP pairs, we adopted the same span-based answer extraction methodology utilized while developing the original QRCD dataset. One Qur'an specialist and two annotators (who are knowledgeable about the Qur'an), extracted the specific answer spans from their respective direct (gold) verse-based answers given by AyaTEC.[8]

[8]Only Qur'an specialists can decide if a verse-based answer represents a *direct* or *indirect* answer to a given question. For a formal definition of a *direct* and *indirect* answer, refer to Malhas and Elsayed (2020).

4.3 Evaluation Setup

In this section, we demonstrate the approach applied to evaluate the performance of participating systems in the MRC task.

4.3.1 Leaderboard and Baseline

As mentioned previously, the leaderboard for the MRC task was hosted on CodaLab (Pavao et al., 2023) with the same conditions over the number of allowed runs. The run file should be in JSON format as in Qur'an QA 2022. However, its format is slightly different. Every answer to each question is a dictionary containing the answer text span, rank, score, start token position, and end token position. The latter two key-value pairs are newly introduced for the task this year.

The baseline for this task is a simple system that gives the whole passage as an answer to the corresponding question. We denote this baseline as *Whole Passage*. Similar to the PR task, we made the baseline code along with submission-checker and evaluation scripts publicly available through the official repository of the shared task.[9]

4.3.2 Evaluation Measures

We chose *partial Average Precision* (pAP) as the main evaluation measure. It is a rank-based measure that integrates *partial matching* to give credit to a QA system that may retrieve an answer that is not necessarily at the first rank and/or *partially* match one of the gold answers (Malhas and Elsayed, 2022). Moreover, pAP is capable of evaluating questions that may have one or more answers in the accompanying passage. This makes pAP more suitable to the MRC task of Qur'an QA 2023 than *partial Reciprocal Rank* (pRR), which was the main evaluation measure for the MRC task in Qur'an QA 2022. Participating systems in the latter task were only required to return *any* answer to a given question even if it has more than one answer in the given passage. Similar to the PR task, the no-answer cases are handled simply by giving full credit to "no answers" system output and zero otherwise. To get an overall evaluation score, the measure is averaged over all questions.

Since the MRC task in Qur'an QA 2023 is different and more challenging than that in Qur'an QA 2022, performance comparisons between the two are not meaningful.

[9] https://gitlab.com/bigirqu/quran-qa-2023

4.4 Results

Twenty nine teams registered for the task. Among these teams, six participated in the final (test) phase with 17 run submissions. The teams are officially ranked based on their best performing submitted run. The performance on the test set of all submitted runs is shown in Table 5, where the runs are ranked by pAP.

It is evident that all teams but one showed superiority over the baseline. The highest pAP score is 0.5711, which was achieved by the TCE (Elkomy and Sarhan, 2023) team. The performance distribution of submitted runs is captured in Figure 5 (in the Appendix). We observed diverse performance across the questions for most of the runs. More details about the teams' approaches are provided next.

Team	Run	pAP
TCE	4dfb8d601	0.5711
TCE	dac0bdf4b	0.5643
Al-Jawaab	tpgp4	0.5457
Al-Jawaab	tgp4	0.5393
TCE	ccc877dca	0.5311
LKAU23	run03	0.5008
LKAU23	run02	0.4989
LowResContextQA	run01	0.4745
LowResContextQA	run02	0.4745
LowResContextQA	run03	0.4745
GYM	run0	0.4613
GYM	ensemble	0.4588
LKAU23	run01	0.4541
GYM	test1	0.4304
Baseline	*WholePassage*	0.3268
PSUT	run2	0.2630
PSUT	RUN3	0.2396
PSUT	RUN1	0.0000

Table 5: MRC evaluation results of all submitted runs ranked by pAP. The team name is removed from the run name to save space. The underlined row is the median run.

4.5 Methods and Analysis

In this section, we provide an overview of the main approaches employed by the 6 participating teams in their submitted runs on the MRC test set. We do that with a focus on the methods employed to address the additional challenges in the MRC task (in its second version); namely, zero-answer questions and multi-answer questions.

Except for Al-Jawaab team (Zekiye and Amroush, 2023) that leveraged generative pre-trained Large Language Models (LLMs) with zero-shot learning setups, all systems of the remaining teams employed encoder-only pre-trained LMs. With the relatively modest size of the QRCDv1.2 dataset, almost all systems leveraged transfer learning by using Arabic pre-trained LMs fine-tuned using large Arabic MRC resources (before fine-tuning using QRCDv1.2) to better perform on the downstream MRC task. Leveraging transfer learning, in the same way, was also heavily witnessed among most of the above-median performing teams in Qur'an QA 2022 (Ahmed et al., 2022; Mostafa and Mohamed, 2022; Wasfey et al., 2022; Premasiri et al., 2022). Interestingly, AraELECTRA (Antoun et al., 2021) and AraBERT (Antoun et al., 2020) LMs maintained their leading performance in both Qur'an QA 2022 and Qur'an QA 2023.

The majority of the systems used one (or more) of the following large Arabic MRC resources for fine-tuning. Ar_TyDiQA (Clark et al., 2020) was used by the TCE (Elkomy and Sarhan, 2023), LowResContextQA (Veeramani and Roy, 2023) and GYM (Mahmoudi and Morshedzadeh, 2023) teams; Arabic SQuADv2.0 (Ahmed, 2023) was used by the GYM and the PSUT teams; and ARCD (Mozannar et al., 2019) and AQQAC (Alqahtani and Atwell, 2018) were used by the LKAU23 (Alnefaie et al., 2023) team. Ensemble and/or self-ensemble learning approaches were also employed by the TCE, LKAU23, LowResContextQA, and GYM teams.

To address the challenge of the zero-answer questions, the TCE, GYM and PSUT teams utilized SQuADv2.0-like fine-tuning (Rajpurkar et al., 2018; Devlin et al., 2019) that uses the [CLS] token to predict the likelihood/probability of a given question to have an answer in the accompanying passage. Interestingly, Al-Jawaab team utilized a carefully hand-crafted prompt (shown in Figure 3) to address the challenge of zero-answer as well as multi-answer questions. The prompt was phrased to instruct their two generative (GPT-4) pre-trained LLMs to answer a given question from its accompanying passage with one *or more* answers, such that they must be extracted from the given passage. The prompt also instructs the model to generate a "no answer" if the given passage does not include an answer to the given question.

As for multi-answer questions, the TCE team employed Maximum Marginal Likelihood (MML) fine-tuning to address this challenge in the MRC task. MML is a form of Bayesian fine-tuning that incorporates uncertainty to preclude the trained systems from being overly confident in a single answer span; thus, distributing the probability among more than one answer span in the accompanying passage of a given question. MML fine-tuning seems to be among the main contributors to the leading performance achieved by TCE (Table 5). No other team addressed this challenge explicitly, other than Al-Jawaab team which used prompt engineering with its generative models (as mentioned above).

```
ءجب على السؤال التالي من النص المرفق فقط .
لا تقم بإضافة أية شرح أو أية إجابة من خارج
النص. اكتب الإجابة أو الإجابات فقط، إن وجدت
أكثر من إجابة اكتبها على شكل تعدادات. الاجابة
يجب أن تكون فقط المقطع أو المقاطع التي تحوي
الجواب بدون أية زيادة. اجعل كل مقطع في سطر
منفصل. إن لم توجد إجابة، اكتب: "No Answer".
```

Figure 3: The handcrafted prompt used by the Al-Jawaab team with their employed generative models.

An important finding by Al-Jawaab team, is that despite their careful prompt engineering scenarios to instruct their generative GPT-4-based models (OpenAI, 2023a; Schreiner, 2023), not to provide out-of-passage answers to a given question, the models sporadically succeeded in providing answer spans strictly from the accompanying passages. Among the main problems was "prompt injection", where parts of the textual prompt instruction/question given to the model are injected back into the generated answer. As such, they applied some post-processing heuristics to the answers obtained by their top performing model.

5 Conclusion

With prevalent *semantic* search approaches on the Holy Qur'an being predominantly ontology-based, we believe that recent neural dense and generative retrieval approaches coupled with the resurgence of the MRC field would pave the way for more intelligent state-of-the-art QA systems on the Holy Qur'an.

To this end, we organized Qur'an QA 2023 shared task, which witnessed the participation of 27 team members from 17 different institutes representing 10 teams. Our shared task in its second version comprised two subtasks; a passage retrieval

(PR) task and a machine reading comprehension (MRC) task. It attracted 9 teams to submit 22 runs for the PR task, and 6 teams to submit 17 runs for the MRC task.

As anticipated, recent transformer-based neural retrieval and reading comprehension approaches were heavily employed by all the participating systems. The majority of the systems deployed encoder-based BERT-like models, whereas generative (decoder-based) GPT-like models were used more sparingly in both tasks. The performance of the systems on the test sets in both tasks indicates that encoder-based transformer models are still taking the lead over generative transformer models. Interestingly, AraELECTRA and AraBERT fine-tuned using large external task-related resources pioneered the Arabic transformers scene. These two models were employed by the best-performing team in each task with self-ensemble. The second best-performing teams in both tasks leveraged generative transformer models (LLMs) using zero-shot learning setups. Though in the PR task, the second ranked team utilized an Arabic-to-English translation module with their retrieval module. The majority of the semantic search/retrieval systems used bi-encoder and cross-encoder architectures independently or jointly. Also, ensemble and self-ensemble approaches were employed by many systems to stabilize prediction fluctuations across runs and/or to enhance prediction accuracy through the wisdom of the crowd.

For zero-answer questions, the best system adopted a thresholding mechanism to identify questions with a low predicted likelihood of having answers in the Holy Qur'an (for Task A), or in the accompanying passage (for Task B). The majority of the teams did not address this challenge *explicitly* in both tasks, other than the second ranked team adopting a naive handcrafted prompt, engineered to instruct their generative GPT-4-based models to return a "no answer" for the MRC task.

As for multi-answer questions in the MRC task, the best performing system employed MML Bayesian fine-tuning to address this challenge. No other team addressed this challenge *explicitly*, other than the second ranked team which used prompt engineering with its generative-based models (as mentioned above). We note that multi-answer (or multi-span) extraction in the literature is an active area of research in the extractive MRC/QA scene that would benefit Qur'anic QA research.

Our prospects towards the third version of the shared task are to aim at including an end-to-end QA task on the Holy Qur'an.

Limitations

The sizes of the AyaTEC and QRCD datasets are relatively modest. This is mainly attributed to the sensitivity of dealing with the sacred Holy Qur'an, for which we have adopted a rigorous and strict process for extracting and annotating the verse-based and span-based answers to the questions of the datasets. Nevertheless, we have foreseen the opportunity to leverage transfer learning and/or model adaptation among other state-of-the-art neural approaches to overcome size-related concerns by question answering systems.

Acknowledgements

We would like to thank all the Qur'an specialists who contributed to extracting/annotating the verse-based and span-based answers to the additional test questions in the datasets; especially Dr. Ahmad Shukri, Professor of Tafseer and Qur'anic Sciences at Qatar University, for his scholarly advice throughout the annotation process of the answers extracted from the Holy Qur'an.

References

Basem H. Ahmed, Motaz K. Saad, and Eshrag A. Refaee. 2022. QQATeam at Quran QA 2022: Fine-Tunning Arabic QA Models for Quran QA Task. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Zeyad Ahmed. 2023. Arabic squad v2.0 dataset based on the popular squadv2.0 with unanswered questions for more challenging task. https://huggingface.co/datasets/ZeyadAhmed/Arabic-SQuADv2.0. Accessed: September 28, 2023.

Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak, Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish. 2022. Proceedinsg of the 5th workshop on osact with shared tasks on qur'an qa and fine-grained hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*.

Hessa A. Alawwad, Lujain A. Alawwad, Jamilah Alharbi, and Abdullah I. Alharbi. 2023. AHJL at Qur'an QA 2023 Shared Task: Enhancing Passage

Retrieval using Sentence Transformer and Translation. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Sarah Alnefaie, Abdullah N. Alsaleh, Eric Atwell, Mohammad Ammar Alsalka, and Abdulrahman Altahhan. 2023. LKAU23 at Qur'an QA 2023: Using Transformer Models for Retrieving Passages and Finding Answers to Questions from the Qur'an. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Mohammad Alqahtani and Eric Atwell. 2018. Annotated corpus of Arabic al-quran question and answer.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. 2022. Arabic natural language processing for Qur'anic research: a systematic review. *Artificial Intelligence Review*, pages 1–54.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, page 4171–4186. Association for Computational Linguistics.

Mohammed Alaa Elkomy and Amany Sarhan. 2023. TCE at Qur'an QA 2023 Shared Task: Low Resource Enhanced Transformer-based Ensemble Approach for Qur'anic QA. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Jimmy Lin and Boris Katz. 2006. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology*, 57(7):851–861.

Ghazaleh Mahmoudi and Yeganeh Morshedzadeh. 2023. GYM at Qur'an QA 2023 Shared Task: Multi-Task Transfer Learning for Quranic Passage Retrieval and Question Answering with Large Language Models. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Rana Malhas. 2023. *Arabic Question Answering on the Holy Qur'an*. Ph.D. thesis, Qatar University.

Rana Malhas and Tamer Elsayed. 2020. AyaTEC: Building a Reusable Verse-based Test Collection for Arabic Question Answering on the Holy Qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(6).

Rana Malhas and Tamer Elsayed. 2022. Arabic Machine Reading Comprehension on the Holy Qur'an using CL-AraBERT. *Information Processing & Management*, 59(6):103068.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 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)*, pages 79–87.

Aly Mostafa and Omar Mohamed. 2022. GOF at Qur'an QA 2022: Towards an Efficient Question Answering For The Holy Qu'ran In The Arabic Language Using Deep Learning-Based Approach. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, page 108–118. Association for Computational Linguistics.

OpenAI. 2023a. GPT-4 Technical Report. Technical report.

OpenAI. 2023b. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model. Accessed: September 27, 2023.

OpenAI. 2023c. OpenAI Embeddings API. https://platform.openai.com/docs/guides/embeddings/. Accessed: September 27, 2023.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, Ruslan Mitkov, Jamal Berrich, and Toumi Bouchentouf. 2022. DTW at Qur'an QA 2022: Utilising Transfer Learning with Transformers for Question Answering in a Low-resource Domain . In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Nils Reimers. 2023. SBERT Semantic Search. https://www.sbert.net/examples/applications/semantic-search/README.html. Accessed: September 27, 2023.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [cls] through ranking by generation. *arXiv preprint arXiv:2010.03073*.

Maximilian Schreiner. 2023. Gpt-4 architecture, datasets, costs and more leaked. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/. Accessed: September 28, 2023.

Marwan N. Swar. 2007. *Mushaf Al-Tafseel Al-Mawdoo'ee*. Dar Al-Fajr Al-Islami, Damascus.

Hariram Veeramani and Kaushik Roy. 2023. LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ahmed Wasfey, Eman Elrefai, Muhammad Marwa, and Nawaz Haq. 2022. Stars at Qur'an QA 2022: Building Automatic Extractive Question Answering Systems for the Holy Qur'an with Transformer Models and Releasing a New Dataset. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Abdulrezzak Zekiye and Fadi Amroush. 2023. Aljawaab at Qur'an QA 2023 shared task: Exploring Embeddings and GPT Models for Passage Retrieval and Reading Comprehension. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Appendix

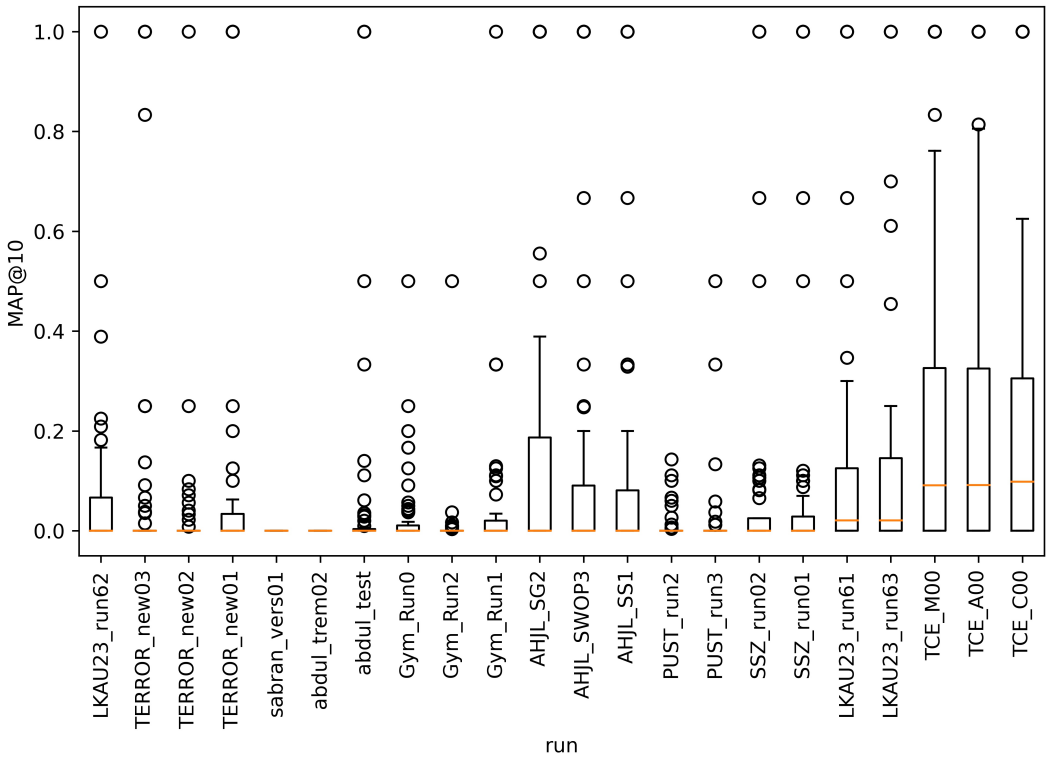

Figure 4: Boxplots for the MAP@10 metric for all submitted runs on the PR task. The plot illustrates the median and inter-quartile distance across questions.

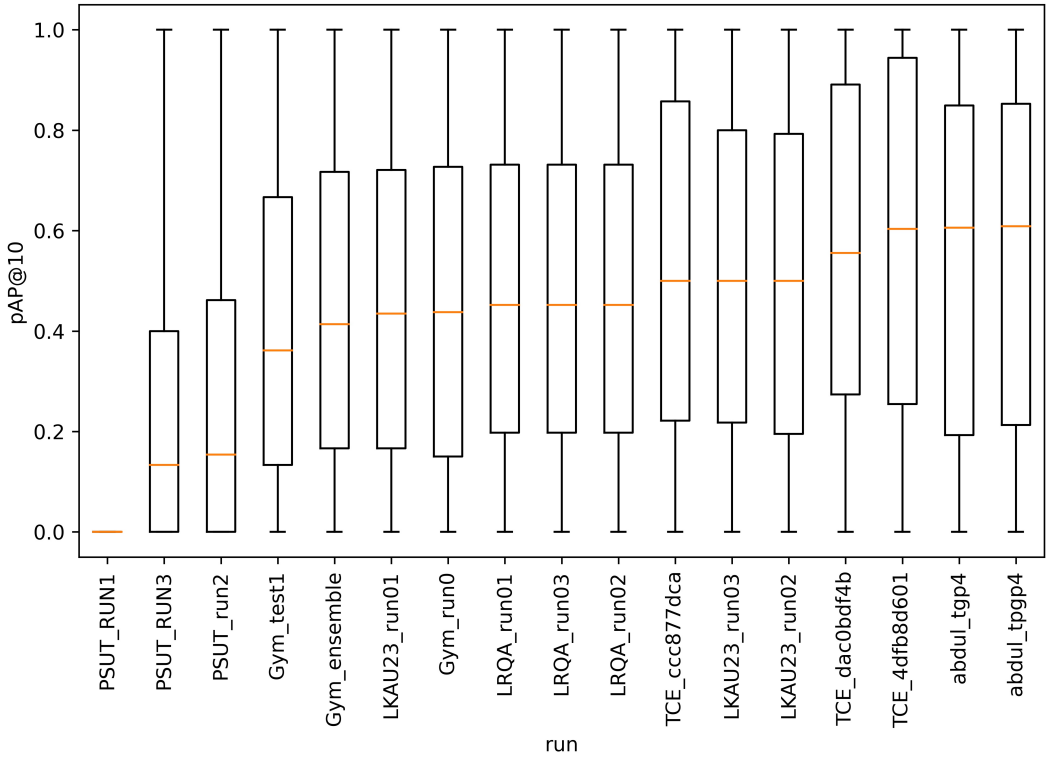

Figure 5: Boxplots for the pAP metric of all submitted runs on task-B. The plot illustrates the median and inter-quartile distance across questions. LRQA is shortened from LowResContextQA.