# Artificial Intelligence for Answering Questions from the Holy Quran and Hadith

UNIVERSITY OF LEEDS

Sarah Saed M Alnefaie

School of Computer Science

University of Leeds

*Submitted in accordance with the requirements for the degree of Doctor of Philosophy*

October 2024

# Dedication

This thesis is dedicated to my mother and to the memory of my beloved father, who passed away before I finished this research. May God have mercy on his soul.

# Declaration

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

During the conduct of this research, nine articles were published as follows:

- Chapter 3:

  Alnefaie, S., Atwell, E., & Alsalka, M. A. Evaluation Criteria of the Islamic Question Answering Systems. IMAN'2022 International Conference on Islamic Applications in Computer Science And Technology Proc.

- Chapter 4:

  1. Alnefaie, S., Atwell, E., & Alsalka, M. A. (2022). Challenges in the Islamic Question Answering Corpora. International Journal on Islamic Applications in Computer Science And Technology, 10(4), 1-10.
  2. Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023). HAQA and QUQA: Constructing two Arabic question–answering Corpora for the Quran and Hadith. In Proceedings of the Conference Recent Advances in Natural Language Processing-Large Language Models for Natural Language Processings (pp. 90-97). INCOMA Ltd., Shoumen, BULGARIA.
  3. Alnefaie, S. S. M., Atwell, E., & Alsalka, M. A. (2023). Using Automatic Question Generation Web Services Tools to Build a Quran Question-and-Answer Dataset. International Journal on Islamic Applications in Computer Science And Technology, 11(2), 1-12.

- Chapter 5:

  Alnefaie, S., Atwell, E., & Alsalka, M. A. (2024, April). Qur'an Passage Ranking Using Transformer Models. In International Conference on Arabic Language Processing (pp. 183-194). Cham: Springer Nature Switzerland.

  This paper was first published in [pp. 183-194, 2024] by Springer Nature and is available online at https://doi.org/10.1007/978-3-031-79164-2_16. Reproduced with permission from Springer Nature.

- Chapter 6:

  1. Alnefaie, S., Alsaleh, A., Atwell, E., Alsalka, M., & Altahhan, A. (2023). LKAU23 at Quran QA 2023: Using Transformer Models for Retrieving Passages and Finding Answers to Questions from the Quran. In Proceedings of ArabicNLP 2023 (pp. 720-727). Association for Computational Linguistics. **(Best Paper Award)**

     **Note**: The question-answering task was carried out by Sarah Alnefaie, the passage retrieval task by Abdullah Al-Saleh, and the remaining authors served as supervisors.

  2. Alnefaie, S., Atwell, E., & Alsalka, M. A. (2024, April). Question Answering over the Arabic Hadith Sharif Using Transformer Models. In International Conference on Arabic Language Processing (pp. 195-206). Cham: Springer Nature Switzerland.

     This paper was first published in [pp. 195-206, 2024] by Springer Nature and is available online at

     https://doi.org/10.1007/978-3-031-79164-2_17. Reproduced with permission from Springer Nature.

- Chapter 7:

  Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023, October). Is GPT-4 a Good Islamic Expert for Answering Quran Questions?. In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023) (pp. 124-133).

- Chapter 8:

  Alnefaie, S., Atwell, E., & Alsalka, M. A. (2024, March). Using the Retrieval-Augmented Generation Technique to Improve the Performance of GPT-4 in Answering Quran Questions. In 2024 6th International Conference on Natural Language Processing (ICNLP) (pp. 377-381). IEEE. (IEEE copyright). It is available online at https://ieeexplore.ieee.org/document/10692797.

# Acknowledgements

I am deeply thankful to Allah for guiding me on the right path and helping me complete this thesis. I also wish to express my heartfelt gratitude to the many individuals who provided encouragement, assistance, and support throughout my studies.

I am deeply grateful to my supervisors, Prof. Eric Atwell and Dr. Mohammad Ammar Alsalka, for their invaluable guidance, constant encouragement, and unwavering support during the preparation of this thesis, which words cannot fully convey.

I offer my deepest thanks to my parents for their unconditional love, constant support, and prayers. I also want to express my love and gratitude to my brother Nawaf and my sisters Meshale, Basma and Maha, who are the "beats of my heart" and my greatest source of inspiration.

I would like to sincerely thank my friends, Nofa, Rana, and Sana, for their support and encouragement.

I am deeply thankful to my country, the Kingdom of Saudi Arabia, for granting me a scholarship to pursue my studies in the UK, and words cannot fully convey my gratitude.

I acknowledge the use of Generative Pre-trained Transformer 4 (GPT-4) in the proof-reading of the thesis before submission.

# Abstract

The aim of this thesis is to develop a system that uses artificial intelligence (AI) models to answer questions related to the Holy Quran and Hadith in classical Arabic (CA). The Holy Quran and Hadith are the two main sources of the Islamic religion.

To achieve this goal, two question-and-answer corpora were developed: one for the Holy Quran, named the Quran Question-Answer (QUQA), and another for Hadith , called the Hadith Question-Answer (HAQA). QUQA is an Arabic dataset focused on the Holy Quran, comprising 3,369 records and over 301,000 tokens. Since some questions may have multiple answers, there are a total of 2,189 unique questions. The verses in the answers represent nearly 47% of the Quran. In the Arabic HAQA dataset for Hadith , there are 1,598 records, over 290,000 tokens, and 1,366 questions.

After creating the datasets, various deep learning (DL) models were explored to obtain answers to the Arabic religious questions. These deep learning models are categorized into pre-trained language models (PLMs) and large language models (LLMs). PLMs, such as BERT, are smaller LLMs that have been fine-tuned for specific downstream tasks. In contrast, LLMs, such as GPT-4, can execute tasks without requiring tailored training data.

When building a question-answering system using PLMs, the system comprises two tasks: passage retrieval (PR) and machine reading comprehension (MRC). In the PR task, the entire Quran or Hadith book is divided into paragraphs. These paragraphs, along with the question, serve as inputs to the model, which retrieves the paragraph containing the answer. There are several approaches to performing the PR task. The method that achieved the best results with the Quran is the Dense Representations Approach (DPR) using the

AraBERT Base model. For the Hadith dataset, two models achieved good results: the hybrid approach using the CAMeL-BERT model and the relevance classification approach using the CAMeL-BERT model. In the MRC task, the inputs are the paragraph selected from the first task and the question. The model identifies the specific answer within the paragraph. I evaluated the performance of approximately nine different Arabic models based on BERT with the Quran and Hadith, employing various methods to improve performance and different datasets for training. Combining the AraBERT Large and AraBERT Base achieved the highest results for the Quran, while CAMeL-BERT achieved the best results for the Hadith.

I also evaluated the effectiveness of LLMs (such as GPT-4) in answering questions related to the Quran and Hadith; however, the outcomes were unsatisfactory. Consequently, I implemented the Retrieval Augmented Generation (RAG) technique, which significantly enhanced the results for Quran-related answers. Nonetheless, these models require considerable further development to achieve a level of understanding comparable to that of humans.

# Abbreviations

| | |
|---|---|
| AGQ | Automatically Generating Questions |
| AI | Artificial Intelligence |
| AQQAC | The Annotated Corpus of Arabic Al-Quran Question and Answer dataset |
| ARBERT | Arabic BERT |
| ARCD | Arabic Reading Comprehension Dataset |
| ARCD | The Arabic Reading Comprehension Dataset |
| BM25 | Best Match 25 |
| CA | Classical Arabic |
| CSVs | Comma-separated values |
| DA | Dialects Arabic |
| DL | Deep Learning |
| DPR | Dense Representations Approach |
| EIAD | English Islamic Articles Dataset |
| ELECTRA | efficiently learning an encoder that classifies token replacements accurately |
| EM | exact match |
| GPT | Generative Pre-trained Transformer Model |
| HAQA | Hadith question–answer dataset |
| HMRC | Hadith machine reading comprehension dataset |
| JISC | The Joint Information Systems Committee |
| LLMs | Large Language Models |
| MADA | the Morphological Analysis and Disambiguation for Arabic toolkit |

| | |
|---|---|
| MAP | Mean Average Precision |
| MARBERT | masked Arabic BERT |
| MCQs | multiple-choice questions |
| MedQuAD | Medical Question Answering Dataset |
| MIMIC-III | Multiparameter Intelligent Monitoring in Intensive Care III Dataset |
| ML | Machine Learning |
| MRC | Machine Reading Comprehension |
| MRR | Mean Reciprocal Rank |
| MSA | Modern Standard Arabic |
| MSMARCO | Microsoft Machine Reading Comprehension dataset |
| NER | named entity recognition |
| NLP | Natural Language Processing |
| NLTK | Python Natural Language Toolkit |
| paP | partial average precision |
| PLMs | Pre-trained Language Models |
| POS | Part of Speech |
| PR | Passage Retrieval |
| pRR | partial reciprocal ranking |
| QA | Questions Answering |
| QARiB | QCRI Arabic and dialectical BERT |
| QG | Question Generation |
| QMRC | Quran MRC dataset |
| QUQA | Quran question–answer dataset |
| R@1 | The Recall of the Top-Ranked Passage |
| RAG | Retrieval-Augmented Generation |
| SQuAD | Stanford Question Answering Dataset |
| SVM | support vector machine |
| T5 | Text-to-Text Transfer Transformer |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| TREC-CAR | TREC-Complex Answer Retrieval dataset |
| TyDiQA-GoldP | Typologically Diverse Question Answering – Gold Passage Dataset |

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Aims and Motivation

Artificial intelligence (AI) is a field of computer science that is currently attracting global attention. It focuses on building machines or systems that can solve problems and perform tasks as humans do by exhibiting similar intelligence. Natural language processing (NLP) is a branch of AI that aims to understand human language (Vajjala *et al.*, 2020). An abundance of resources, such as computing power and data, has led to major advances in NLP (Goodfellow *et al.*, 2016). These developments rely on using machine learning (ML) and deep learning (DL) algorithms and models to solve NLP tasks. These algorithms and models require a vast amount of data, including examples from which to learn and significant processing power to perform tasks automatically.

NLP has been applied across various fields to automate numerous tasks. For instance, in the medical field, it has been used to develop diagnostic models capable of detecting chronic diseases in their early stages (Locke *et al.*, 2021). Sentiment analysis, an NLP task utilised in the business sector, helps companies automatically understand consumer feedback on products and gain deeper insights into consumer needs (Solairaj *et al.*, 2023). The modelling of Islamic religious texts is an intriguing area within NLP, as these texts are sacred and serve as essential sources of knowledge for both Muslim and non-Muslim knowledge seekers (Atwell *et al.*, 2011; Azmi *et al.*, 2019). Successful modelling of these texts assists those interested in extracting information from them.

There is a notable shortage of available Islamic datasets that can be utilised for training and evaluating DL models and ML algorithms (Bounhas, 2019; Malhas & Elsayed, 2020). These constraints hinder the effective application of these models and algorithms in research involving Islamic texts. To address this issue, my goal was to enhance Arabic Islamic language resources by creating two datasets: one consisting of question-and-answer pairs for the Quran and the other for the Hadith. The Holy Quran and the Hadith are primary sources for millions of Muslims around the globe. They provide essential guidance on legislation, knowledge, wisdom, teachings and a comprehensive understanding of the religion, making them invaluable resources for answering pertinent questions (Atwell *et al.*, 2010).

Currently, only two question–answer datasets are available for the Holy Quran: the Annotated Corpus of Arabic Al-Quran Question and Answer (AQQAC) (Alqahtani, 2019) and AyaTEC (Malhas & Elsayed, 2020). These datasets are relatively small in size. This limitation during the training phase has impacted performance in their use with pre-trained transformer language models (Malhas & Elsayed, 2022; Wasfey *et al.*, 2022). Additionally, the limited size of an evaluation dataset hinders the ability to assess the performance of systems with depth and accuracy. Consequently, there is a need to create a larger question–answer dataset for the Quran.

There is currently no Hadith question–answer dataset available to the research community, despite many religious questions and answers about the Hadith being scattered across websites and books in various formats. The absence of this corpus has resulted in the lack of a standardised dataset for evaluating the performance of question-answering (QA) systems in the Hadith domain. Consequently, researchers have had to create their own datasets to assess the performance of their systems by collecting question–answer pairs from various websites and books, or generating them independently (Abdi *et al.*, 2020; Maraoui *et al.*, 2021; Neamah & Saad, 2017). As a result, each of these corresponding systems utilises different dataset sizes and types of question–answer pairs, making it difficult to compare the system performance fairly. Therefore, building a question–answer dataset for the Hadith is essential.

In addition to a research question on the development of these two datasets, other research questions have been created that pertain to investigating the performance of pre-trained language models (PLMs) and large language models (LLMs) in finding answers to questions in the Quran and Hadith. Developing a QA system for the Quran and Hadith is important because 1.9 billion Muslims around the world rely on these texts for religious teachings to guide their daily lives, and these resources contain a wealth of scientific knowledge and wisdom (Malhas, 2023). Additionally, many non-Muslims seeking knowledge require answers to their questions and could also benefit from such a QA system . A recent systematic review on Arabic NLP for Quranic texts identified the development of intelligent systems for the Quran as one of the open issues for future research (Bashir *et al.*, 2023).

The Arabic language includes variants such as Modern Standard Arabic (MSA), Dialectal Arabic (DA) and Classical Arabic (CA). In this thesis, the texts of the Quran and Hadith were studied in CA because AI models and algorithms have not been extensively explored for their performance in this language (Guellil *et al.*, 2021). Many research studies have focused on the MSA variant, while fewer have addressed the CA variant (Malhas, 2023).

The Arabic language, in general, is considered challenging for NLP for several reasons. For example, there are no capital letters, and a single term may have many morphemes, as shown in Figure 1.1.

There is also a challenge in CA related to diacritics. Figure1.2 shows text with and without diacritics to illustrate the difference. Diacritics are significant because they clarify the meaning of a word, and the meaning changes as the diacritics change, as shown in Figure1.3. In the preprocessing stage of most NLP tasks, diacritics are often removed from the Quran and Hadith texts. This step adds ambiguity to the meanings of words, increasing the challenge for NLP systems (Bashir *et al.*, 2023; Malhas, 2023).

## 1.2 Research Questions

1. **RQ1:** Is it possible to develop two question–answer datasets, one for the Quran and the other for the Hadith, and make them available to the NLP

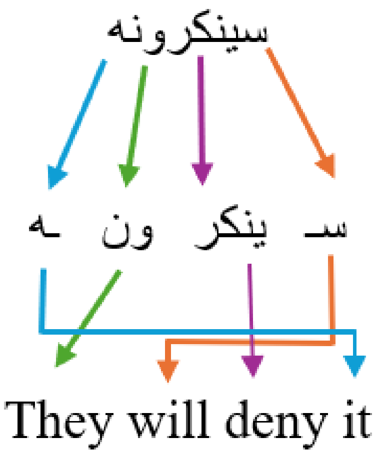



Figure 1.1: An example of many morphemes in a single Arabic term.

Text with diacritics: اَلْحَمْدُ لِلّٰهِ رَبِّ الْعَالَمِيْنَ

Text without diacritics: الحمد لله رب العالمين

Figure 1.2: An example of text with and text without diacritics.

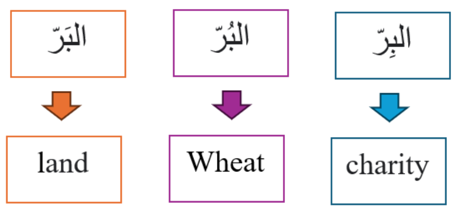



Figure 1.3: An example of the effect of diacritics on the meaning of a word.

research community?

2. **RQ2:** Can PLMs answer questions from the Quran and Hadith in Arabic?
3. **RQ3:** Does the size of the new Quran question–answer dataset affect the performance of the PLMs?
4. **RQ4:** Are LLMs, such as GPT-4, effective Islamic experts for answering Quran questions?
5. **RQ5:** Does using the retrieval-augmented generation (RAG) technique improve the performance of GPT-4 in answering Quranic questions in Arabic?

## 1.3 Contributions

The novel contributions of this thesis are described below.

### 1.3.1 Quran Question–Answer Corpus and Hadith Question–Answer Corpus

The first contribution of this thesis is the creation of a large-scale question–answer dataset for the Quran, which includes challenging questions that cover a significant portion of the Holy Quran. This resource is called the Quran Question–Answer (QUQA) corpus. The name is an abbreviation formed from the first two letters of the word 'Quran', the first letter of the word 'question' and the first letter of the word 'answer'.

The second contribution is the Hadith Question–Answer (HAQA) corpus. This dataset consists of questions and answers derived from the Hadith. Similar to the naming convention of the Quran dataset, the Hadith dataset's name is an abbreviation composed of the first two letters of the word 'Hadith', the first letter of the word 'question' and the first letter of the word 'answer'.

Two question–answer datasets for the Quran are available; however, they have several drawbacks, including their small size. The two datasets were provided in different formats and structures. First, I reviewed the two collections and selected questions from them that included answers from the Holy Quran. I then chose a unified structure for a new dataset and merged the selections from the two original datasets into one file. Finally, I expanded it using additional sources. To my knowledge, no Hadith dataset is available to the research community. These two resources will be beneficial to researchers, serving as gold standard datasets for evaluating QA systems. When researchers access these datasets and use them to evaluate their systems, they facilitate comparison of the results. They can also be utilised to train AI models. Both corpora are in CA, which is considered a low-resource language. They are available on GitHub [1]. Chapter 4 presents the methodology for constructing these two datasets.

### 1.3.2 Evaluating Pre-trained Language Models in Answering Questions from the Holy Quran and the Hadith

Upon completing the two datasets, I studied the application of transformer models in answering questions from the Holy Quran and Hadith through various experiments. This involved two key tasks: a passage retrieval (PR) task and a machine reading comprehension (MRC) task, both of which were performed using the QUQA and HAQA datasets. In the PR task, the text from which the answer is to be extracted, such as the Holy Quran, is divided into passages. When presented with a free-text question in MSA and a collection of passages from the Holy Quran, a system is needed to generate a ranked list of passages that may contain the answer(s) to the specified question. The passage that receives the highest score from the PR task, indicating a high probability of containing the answer, is then sent to the MRC task. In the MRC task, the inputs to the model are the text passage and the question. The model extracts the correct answer, and its output is a ranked list of answers.

I evaluated and compared all the models available for answering questions from the Holy Quran and Hadith in the PR and MRC tasks. The results of the PR task are presented and discussed in Chapter 5, while the MRC task is presented and discussed in Chapter 6. Some previous researchers have developed models for Quranic PR; however, the performance of the models was poor, the researchers did not validate the performance of all the existing approaches and

[1] https://github.com/scsaln/HAQA-and-QUQA

they all used small datasets for training (Alnefaie *et al.*, 2023a; Mahmoudi *et al.*, 2023; Zekiye & Amroush, 2023). In the MRC task, extant research has examined the use of models to find answers to questions in the Quran (Malhas & Elsayed, 2022). However, these studies also utilised small datasets, which negatively affected performance. In this thesis, I used the QUQA dataset, which is relatively large. To the best of my knowledge, no previous study has analysed PR and MRC tasks for Hadith texts.

### 1.3.3 Assessing Large Language Models in Answering Questions from the Holy Quran and the Hadith

Recently, most research has focused on investigating the performance of LLMs to identify their behaviour, defects and weaknesses to contribute to their development. This thesis explored the performance of these models in answering Quranic and Hadith questions in Arabic. The results of this study are presented in Chapter 7. This research is the first to investigate the performance of GPT-4 with religious texts in Arabic. RAG is the latest optimisation method used in various domains (Katz *et al.*, 2024; Kung *et al.*, 2023; Wood *et al.*, 2023). In this thesis research, I applied this technique in the Islamic domain to answer questions from the Quran. Details of this technique and its results can be found in Chapter 8. Verifying the performance of these models will help researchers, interested individuals, and religious scholars determine whether the models can be relied on in the future.

## 1.4 Thesis Outline

As shown in Figure 1.4, this thesis is divided into four sections. The first section consists of three chapters. Chapter 1 presents an introduction to the thesis topic. Chapter 2 provides background on two important subjects: (1) the Arabic language, including its types, features and related challenges and (2) information about the two most important texts in the Islamic religion, the Holy Quran and Hadith. Chapter 3 presents a literature review that offers a general and comprehensive overview of QA systems.

The second section primarily focuses on the approach used to construct the two datasets, QUQA and HAQA. Chapter 4 begins with a literature review of the datasets and the various methods employed to construct them. It then presents the methodology for building the two corpora and analyses the resulting datasets.

The third section reviews the evaluation of AI models. Chapter 5 focuses on exploring the performance of PLMs in the PR task for Quran questions and Hadith questions. Chapter 6 investigates the performance of nine transformer models in the MRC task. Chapter 7 utilises one LLM, GPT-4, to find answers to these questions. Chapter 8 discusses the use of RAG technology to enhance the performance of the LLM.

The last section contains only Chapter 9, which reviews the conclusion, limitations of the work and possible directions for future research.

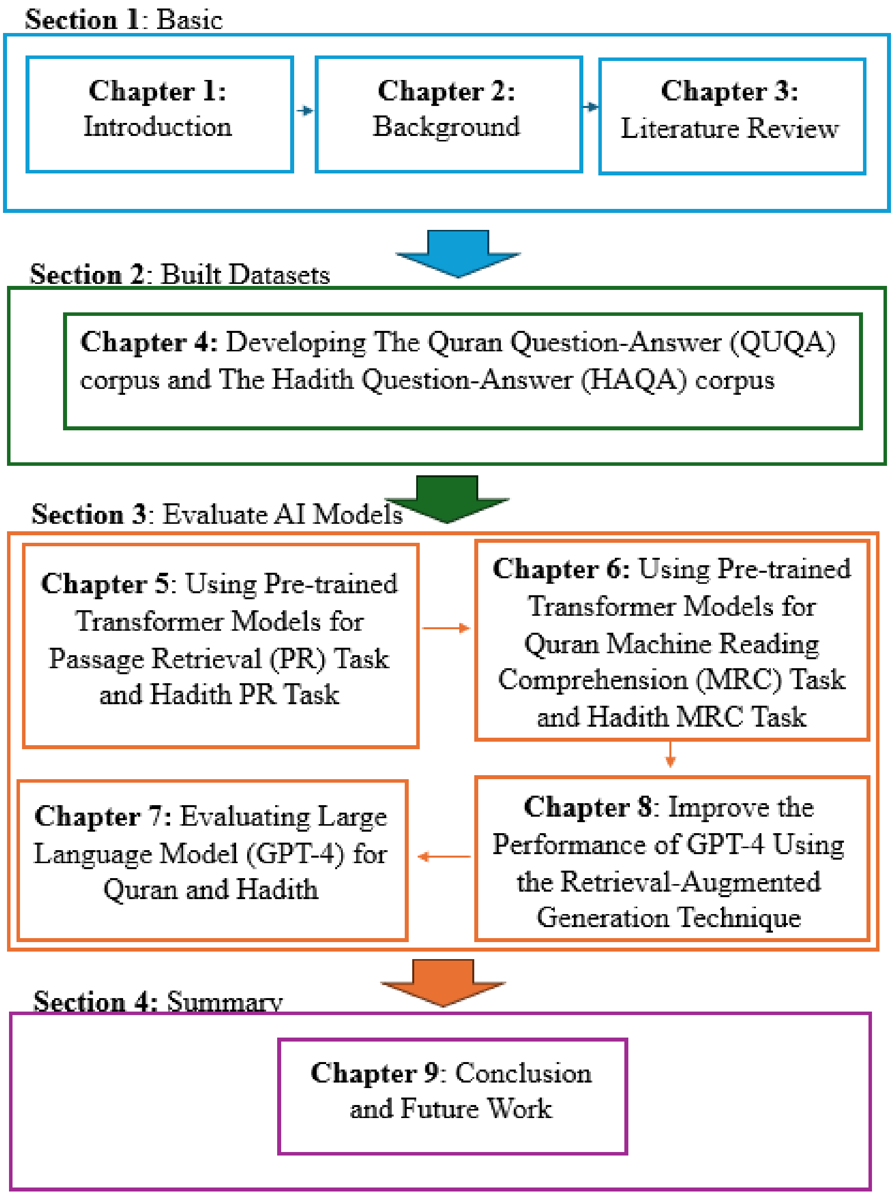



Figure 1.4: Thesis Outline.

# Chapter 2

# Islamic and Linguistic Background

## 2.1 Introduction

This chapter delves into the linguistic and historical knowledge of Islamic religious texts, particularly the Quran and Hadith, and discusses the challenges related to using them as resources. Globally, these texts are crucial for Muslims but can be difficult for individuals to grasp and interpret due to language issues. This chapter explores the obstacles associated with these valuable resources for deeper exploration. In addition, it provides a concise overview of the Arabic language.

## 2.2 Islamic books

There are five foundational books for learning about the Islamic religion: the Holy Quran, Hadith, Aqeedah, Fiqh and Quran Exegesis (Tafsir) (Al-Munajjid, 2001). I discuss the Quran and Hadith in detail in the following sections of this chapter. The books of Aqeedah address what a person believes about himself and his spirit with complete certainty. In Arabic, Fiqh means knowledge; however, the term 'fiqh' in a Shar'i context refers to 'the knowledge of practical, minor Shar'i rulings that are derived from detailed evidence and proof' (Al-Munajjid, 2021). The goal of Quran exegesis (Tafsir) is to understand the meanings of the Holy Quran and what God Almighty wants from his creation. Therefore, many scholars are interested in the science of interpretation. The process of interpreting the Quran began during the era of prophecy. The Prophet Muhammad, may God bless him and grant him peace, is considered the first authority in interpreting the Quran. He interpreted the verses using two methods. In the first method, he interpreted Quranic verses with references to other Quranic verses. In the second method, he interpreted the verses through his words and actions (Ahmed AbdulKadhim Tellab, 2023; Muhammad, 2004).

An example of the first method is that the meaning of a word in a Quranic verse is explained by utilising another Quranic verse, as shown in Figure 2.1. For example, the term 'hard clay', 'سِجِّيلٍ', which appears in Surah Al-Hijr, verse 74, has multiple meanings. The specific meaning of 'hard clay', 'سِجِّيلٍ', as intended in this Quranic verse is clarified in Surah Adh-Dhariyat, verse 33, where it is referred to as clay, 'طِينٍ' (bin Jarir al Tabari, 2001).

**The Holy Quran [Sura: Al-Hijr (15): Verse 74]:**

فَجَعَلْنَا عَالِيَهَا سَافِلَهَا وَأَمْطَرْنَا عَلَيْهِمْ حِجَارَةً مِنْ سِجِّيلٍ ﴿74﴾

And we made the highest part [of the city] its lowest and rained upon them stones of hard clay. (74)

**The Holy Quran [Sura: Adh-Dhariyat (51): Verse 33]:**

لِنُرْسِلَ عَلَيْهِمْ حِجَارَةً مِنْ طِينٍ ﴿33﴾

To send down upon them stones of clay (33)

Figure 2.1: An example of an explanation of the Quran verse by another the Quran verse.

Sometimes, the wording of a Quranic verse is ambiguous and requires clarification and additional detail, and no other Quranic verse specifically defines this ambiguity, such as the word 'more' in verse 26 of Surah Yunus, as shown in Figure 2.2. Therefore, I refer to the Hadith to understand its meaning, as there are many hadiths in which the Prophet, may God bless him and grant him peace, has detailed and explained their meanings. In particular, the Prophet, may God bless him and grant him peace, clarified the word ‘more’ by looking at God Almighty (bin Jamil Zaino, 1997).

**The Holy Quran [Sura: Yunus (10): Verse 26]:**

لِّلَّذِينَ أَحْسَنُوا الْحُسْنَىٰ وَزِيَادَةٌ ۖ وَلَا يَرْهَقُ وُجُوهَهُمْ قَتَرٌ وَلَا ذِلَّةٌ ۚ أُولَٰئِكَ أَصْحَابُ الْجَنَّةِ ۖ هُمْ فِيهَا خَالِدُونَ ﴿26﴾

For them who have done good is the best [reward] and extra. No darkness will cover their faces, nor humiliation. Those are companions of Paradise; they will abide therein eternally (26)

**The Hadith:**

حَدَّثَنَا مُحَمَّدُ بْنُ بَشَّارٍ، حَدَّثَنَا عَبْدُ الرَّحْمَنِ بْنُ مَهْدِيٍّ، حَدَّثَنَا حَمَّادُ بْنُ سَلَمَةَ، عَنْ ثَابِتٍ الْبُنَانِيِّ، عَنْ عَبْدِ الرَّحْمَنِ بْنِ أَبِي لَيْلَى، عَنْ صُهَيْبٍ، عَنِ النَّبِيِّ صلى الله عليه وسلم فِي قَوْلِهِ: (لِلَّذِينَ أَحْسَنُوا الْحُسْنَى وَزِيَادَةٌ ) قَالَ " إِذَا دَخَلَ أَهْلُ الْجَنَّةِ الْجَنَّةَ نَادَى مُنَادٍ إِنَّ لَكُمْ عِنْدَ اللَّهِ مَوْعِدًا . قَالُوا أَلَمْ يُبَيِّضْ وُجُوهَنَا وَيُنَجِّنَا مِنَ النَّارِ وَيُدْخِلْنَا الْجَنَّةَ قَالُوا بَلَى . قَالَ فَيُكْشَفُ الْحِجَابُ قَالَ فَوَاللَّهِ مَا أَعْطَاهُمْ شَيْئًا أَحَبَّ إِلَيْهِمْ مِنَ النَّظَرِ إِلَيْهِ " . قَالَ أَبُو عِيسَى هَذَا حَدِيثٌ إِنَّمَا أَسْنَدَهُ حَمَّادُ بْنُ سَلَمَةَ وَرَفَعَهُ . وَرَوَى سُلَيْمَانُ بْنُ الْمُغِيرَةِ وَحَمَّادُ بْنُ زَيْدٍ هَذَا الْحَدِيثَ عَنْ ثَابِتٍ الْبُنَانِيِّ عَنْ عَبْدِ الرَّحْمَنِ بْنِ أَبِي لَيْلَى قَوْلَهُ .

For those who do good is the best (reward) and even more- the Prophet (s.a.w) said: "When the people of Paradise enter Paradise, a caller shall call out: 'Indeed you have a promise with Allah.' They will say: 'Did he not whiten our faces, save us from the Fire, and admit us into Paradise?' They will say: 'Indeed.' Then the Veil shall be lifted." He said: "So, by Allah, He did not grant them anything more beloved to them than looking at Him."

Figure 2.2: An example of an explanation of the Quran verse by Hadith.

## 2.3 The Holy Quran

Muslims define the Quran as ‘the speech of Allah Almighty, which He revealed to His Messenger Muhammad (may Allah’s peace and blessings be upon him), and His servants worship Him by reciting it. It is written in Mus-hafs and has reached us through successive transmission’. The phrase ‘worshipped by its recitation’ means that the believer draws closer to God by reading the Quran. The term ‘successive transmission’ refers to the way it has been passed down from one group of people to another group of people over centuries, making it impossible for them to have agreed to lie. It features miraculous Arabic pronunciation and embodies a miracle of Muhammad (peace be upon him) due to its news of previous nations, scientific miracles and eloquence. Among the eloquence of the Quran is the fact that the Messenger Muhammad, may God’s prayers and peace be upon him, challenged his people to produce a complete Quran, as stated in the verse shown in Figure 2.3. Figure 2.4 illustrates the verse in which, when his people could not produce a full Quran, he challenged them to create ten surahs, but they could not. He then reduced the challenge to just one surah, and they also failed, as explained in Figure 2.5.

**The Holy Quran [Sura: Al-Isra (17): Verse 88]:**

قُلْ لَئِنِ اجْتَمَعَتِ الْإِنْسُ وَالْجِنُّ عَلَىٰ أَنْ يَأْتُوا بِمِثْلِ هَٰذَا الْقُرْآنِ لَا يَأْتُونَ بِمِثْلِهِ وَلَوْ كَانَ بَعْضُهُمْ لِبَعْضٍ ظَهِيرًا ﴿88﴾

Say, "If mankind and the jinn gathered in order to produce the like of this Qur'an, they could not produce the like of it, even if they were to each other assistants." (88)

Figure 2.3: The verse in which the messenger challenged his people to produce a complete Quran.

The Quran was revealed to Muhammad, may God’s blessings and peace be upon him, through Gabriel (peace be upon him). It begins with Surah Al-Fatihah and concludes with Surah Al-Nas. Its revelation took place over 23 years, approximately 1,445 years ago. The language of the Quran is CA. Two billion Muslims around the world believe in the Quran. Muslims must understand it correctly, deeply and accurately, as they derive teachings that affect all aspects of their lives from it (Alqahtani & Atwell, 2018; Altammami & Atwell, 2022; Brown, 2017; Mustafa Deeb Al-Bagha, 1998).

**The Holy Quran [Sura: Hud (11): Verse 13]:**

أَمْ يَقُولُونَ افْتَرَاهُ ۖ قُلْ فَأْتُوا بِعَشْرِ سُوَرٍ مِثْلِهِ مُفْتَرَيَاتٍ وَادْعُوا مَنِ اسْتَطَعْتُمْ مِنْ دُونِ اللَّهِ إِنْ كُنْتُمْ صَادِقِينَ ﴿13﴾

Or do they say, "He invented it"? Say, "Then bring ten surahs like it that have been invented and call upon [for assistance] whomever you can besides Allah, if you should be truthful." (13)

Figure 2.4: The verse in which the messenger challenged his people to produce ten suras like the suras of the Quran.

**The Holy Quran [Sura: Al-Baqara (2): Verse 23]:**

وَإِنْ كُنْتُمْ فِي رَيْبٍ مِمَّا نَزَّلْنَا عَلَىٰ عَبْدِنَا فَأْتُوا بِسُورَةٍ مِنْ مِثْلِهِ وَادْعُوا شُهَدَاءَكُمْ مِنْ دُونِ اللَّهِ إِنْ كُنْتُمْ صَادِقِينَ ﴿23﴾

And if you are in doubt about what We have sent down upon Our Servant [Muhammad], then produce a surah the like thereof and call upon your witnesses other than Allah, if you should be truthful (23)

Figure 2.5: The verse in which the messenger challenged his people to produce only one sura, like the suras of the Quran.

### 2.3.1 Structure of the Holy Quran

The Holy Quran covers many topics: former nations, the unseen, legislation in Islam, stories of prophets, faith in God and believers, jihad, battles and wars, the universe, humanity, scientific knowledge, God's creations, provisions of Islam, worship, wisdom, the Prophet Muhammad, the purpose of creation, the linguistic aspects of the Quran and others. Topics are addressed by verses scattered throughout the Quran, in a series of connected verses or in a single verse. A single verse may cover a range of subjects. This feature is known as 'unstructured topic diversity'. There are 114 chapters in this book, and each chapter is called a surah. A surah is a series of verses (ayahs) of varying lengths. The total number of verses is up to 6,236 (ayahs). The surahs are very long at the beginning of the Quran, with their lengths gradually decreasing. Surah Al-Baqarah, located at the beginning of the Quran, contains 286 verses, while the last Surah, Surat Al-Nas, contains only 6 verses (Alqahtani & Atwell, 2018; Altammami & Atwell, 2022; Brown, 2017). An example of a Surah from the Holy Quran is shown in Figure 2.6.

### 2.3.2 Challenges in the Quranic Text

For centuries, scholars have believed that the Quran could not be translated into other languages because the Quranic text is miraculous. It is very difficult to convey the exact meaning of the verses using other languages; therefore, its translation was delayed for centuries. Dr Mohammad Abdul Hakim Khan wrote the first English translation of the Holy Quran in 1904. After that, many translators published their own versions of the Quran in English, with the goal of competing to convey the meaning more accurately. A group of researchers subsequently reviewed these translations, discovered many errors and analysed the reasons behind these mistakes. In this chapter, I review several of these reported errors. Given the difficulties and challenges translators face in understanding the verses correctly, AI models may encounter similar challenges, leading to incorrect answers to questions (Faqeer, 2017).

The rhetorical, moral, semantic and syntactic characteristics that distinguish the Quranic text make it unique compared to other Arabic works. Its words

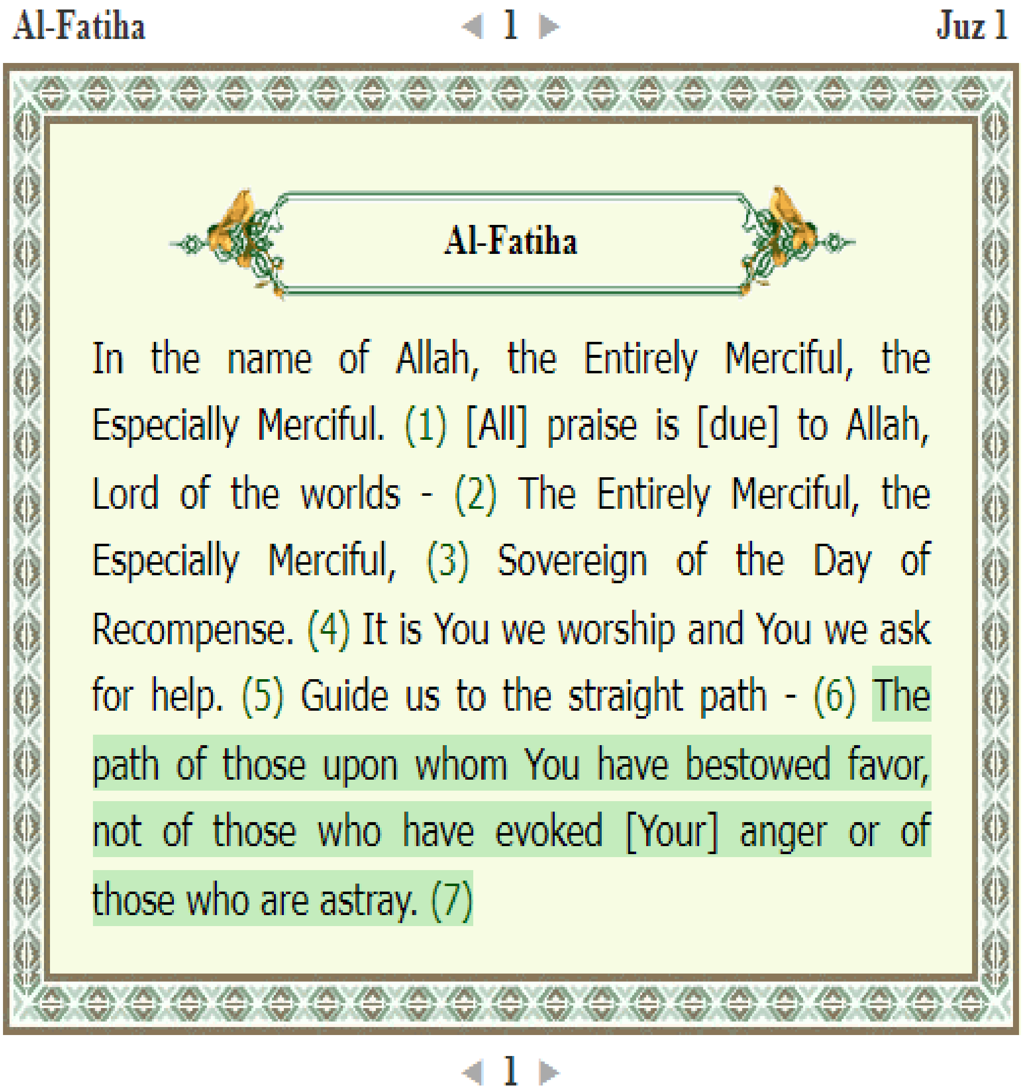



Figure 2.6: An example of a surah from the Holy Quran.

contain layers of embedded meanings, which require religious scholars and dictionaries to understand correctly, deeply and accurately. Additionally, many words have special meanings in the Holy Quran, and, thus, their interpretations in that text differ from how they would be understood when seen in other Arabic works (Abdul-Raof, 2013).

The following points outline some errors or challenges in translations of the Quran into English that researchers have discussed (Siddiek, 2017):

1. **Synthetic-Semantic Ambiguity**

   Many translators believe that a correct translation requires only a literal rendering of words and the proper application of grammatical rules. In regard to the Quran, such a view neglects the text's unique characteristics, which are not found in the English language. Neglecting these characteristics during the translation process leads to an incorrect understanding of the sentences, resulting in ambiguity in the overall meaning and an inaccurate translation of the Quranic text. An example of errors that arise in English translations of the Quran, specifically due to neglecting the flexibility of word order in Arabic sentences, is shown in Figure 2.7.

**The Holy Quran [Sura: Fater (35): Verse 28]:**

... إِنَّمَا يَخْشَى اللَّهَ مِنْ عِبَادِهِ الْعُلَمَاءُ ۗ ... ﴿28﴾

**Translator:** Arberry
**English Translation**: …Only those of His servants fear God who have knowledge…(28)

**Translator:** Pickthal
**English Translation**: … The erudite among His bondsmen fear Allah alone…(28)

**The correct translation**
**English Translation**: … From among His servants, only those who have knowledge fear Allah…(28)

Figure 2.7: Example of Synthetic-Semantic Ambiguity.

In addition, Quranic text relies on diacritical marks to determine the grammatical position of words within a sentence. Therefore, the Quranic text is flexible in determining the placement of words in a sentence. For example, the word ‘Allah-God’ is expected to be the subject because it appears after the verb. However, the diacritical mark on the word is al-fatha; therefore, its correct grammatical position is that of an object. Because diacritics do not exist in English, translators may become confused about their meaning and impact, leading to incorrect translations.

In Arberry’s translation of the Quran into English , his misunderstanding of a sentence led him to link the phrase ‘who has knowledge’ with ‘God’ instead of linking it to ‘His servants.’ In his English translation, Pickthal understood that there was an implicit exception, so he placed the word ‘alone’ in the wrong position. However, the correct word is ‘only’; its placement before the phrase ‘who have knowledge’ is based on the opinions of a number of scholars involved in the interpretation of the Quran (Darwish, 1983).

2. **Lexical Semantic Ambiguity**

In lexical semantic ambiguity, the reader faces difficulty in determining the correct meaning of a word. There are two types of Quranic context: explicit and implicit. In an explicit context, a word’s meaning is clear and directly accessible to the reader. In contrast, the implicit context requires understanding interpretations of the entire verse or surah and its meaning to determine a word’s meaning.

A word may have multiple meanings, and its correct meaning in a text is determined by the context. This phenomenon is called polysemy (Ali *et al.*, 2014). For example, the Arabic word for ‘clothing’ لِبَاسٌ appears in the Holy Quran with several meanings, including mixing one thing with another, as shown in Figure 2.8. Another meaning is serenity, as illustrated in the verse in Figure 2.9. Figure 2.10 shows the verse in which ‘clothing’ refers to faith or good deeds.

**The Holy Quran [Sura: Aal-i-Imran (3): Verse 71]:**

يَا أَهْلَ الْكِتَابِ لِمَ تَلْبِسُونَ الْحَقَّ بِالْبَاطِلِ وَتَكْتُمُونَ الْحَقَّ وَأَنْتُمْ تَعْلَمُونَ ﴿71﴾

**Translator:** Yusuf Ali
**English Translation**: Ye People of the Book! Why do ye clothe Truth with falsehood, and conceal the Truth, while ye have knowledge? (71)

**Translator:** Arberry
**English Translation**: People of the Book! Why do you confound the truth with vanity, and conceal the truth and that wittingly? (71)

**Translator:** Abdel Haleem
**English Translation**: People of the Book, why do you mix truth with falsehood? Why do you hide the truth
when you recognize it? (71)

Figure 2.8: The first meaning of the word clothe.

**The Holy Quran [Sura: Al-Baqara (2): Verse 187]:**

... هُنَّ لِبَاسٌ لَكُمْ وَأَنْتُمْ لِبَاسٌ لَهُنَّ ۗ...﴿187﴾

**Translator:** Yusuf Ali
**English Translation**: …They are your garments and ye are their garments …(187)

**Translator:** Arberry
**English Translation**: …they are a vestment for you, and you are a vestment for them … (187)

**Translator:** Abdel Haleem
**English Translation**: … they are [close] as garments to you, as you are to them … (187)

Figure 2.9: The second meaning of the word clothe.

**The Holy Quran [Sura: Al-A'raf (7): Verse 26]:**

... وَلِبَاسُ التَّقْوَىٰ ذَٰلِكَ خَيْرٌ ۚ...﴿26﴾

**Translator:** Yusuf Ali
**English Translation**: … But the raiment of righteousness, - that is the best…(26)

**Translator:** Arberry
**English Translation**: … ; and the garment of godfearing -- that is better… (26)

**Translator:** Abdel Haleem
**English Translation**: … The garment of Godconsciousness is the best of all garments… (26)

Figure 2.10: The second meaning of the word clothe.

Sometimes, translators face difficulty in finding an appropriate word in English for the meaning in the original. This section discusses two issues related to this topic. First, in some languages, there may be no single word that represents a word in another language while expressing the same meaning correctly and accurately. Second, it is common in some languages to link two or more words in an idiomatic phrasing, but doing so in another language may be unnatural and may not convey the same meaning.

Because of similar experiences and their shared humanity, people have created similar words in various languages, especially for simpler objects and ideas. However, the different environments in which various languages have developed and a multitude of complex factors, such as a variety of synonyms for words in most languages, have made each vocabulary fairly unique. This makes it difficult to find words in one language that match words in another language with the same meaning. For example, consider the following three words in Arabic: نَقِيرًا, قِطْمِيرٍ and فَتِيلًا. They refer to different parts of dates. One English translator could not find literal translations for these terms, so he translated them all figuratively into the general word ‘date,’ as shown in Figure 2.11 (Altammami, 2023b).

**The Holy Quran [Sura: An-Nisa (4): Verse 49]:**

أَلَمْ تَرَ إِلَى الَّذِينَ يُزَكُّونَ أَنْفُسَهُمْ ۚ بَلِ اللَّهُ يُزَكِّي مَنْ يَشَاءُ وَلَا يُظْلَمُونَ فَتِيلًا ﴿49﴾

**Translator:** Pickthalls
**English Translation**: Hast thou not seen those who praise themselves for purity? Nay, Allah purifieth whom He will, and they will not be wronged even the hair upon a date-stone (49).

**The Holy Quran [Sura: An-Nisa (4): Verse 124]:**

وَمَنْ يَعْمَلْ مِنَ الصَّالِحَاتِ مِنْ ذَكَرٍ أَوْ أُنْثَىٰ وَهُوَ مُؤْمِنٌ فَأُولَٰئِكَ يَدْخُلُونَ الْجَنَّةَ وَلَا يُظْلَمُونَ
نَقِيرًا ﴿124﴾

**Translator:** Pickthalls
**English Translation**: And whoso doeth good works, whether of male or female, and he (or she) is a believer, such will enter paradise and they will not be wronged the dint in a date-stone (124).

**The Holy Quran [Sura: Fatir (35): Verse 13]:**

يُولِجُ اللَّيْلَ فِي النَّهَارِ وَيُولِجُ النَّهَارَ فِي اللَّيْلِ وَسَخَّرَ الشَّمْسَ وَالْقَمَرَ كُلٌّ يَجْرِي لِأَجَلٍ مُسَمًّى ۚ
ذَٰلِكُمُ اللَّهُ رَبُّكُمْ لَهُ الْمُلْكُ ۚ وَالَّذِينَ تَدْعُونَ مِنْ دُونِهِ مَا يَمْلِكُونَ مِنْ قِطْمِيرٍ﴿13﴾

**Translator:** Pickthalls
**English Translation**: He maketh the night to pass into the day and He maketh the day to pass into the night. He hath subdued the sun and moon to service. Each runneth unto an appointed term. Such is Allah, your Lord; His is the Sovereignty; and those unto whom ye pray instead of Him own not so much as the white spot on a date-stone (13).

Figure 2.11: Example of figurative words.

Some words are usually mentioned together to express a certain meaning in one language but are unusual when used in another. Therefore, the meaning will not be conveyed correctly when the sentence is translated literally from Arabic to English. For example, mentioning the two words 'eat' أكل and 'money' أموال together is very common in the Quranic text, as shown in Figure 2.12, where the two words appear next to each other in various verses. However, this connection between the two words typically does not exist in English. English translators of the Quran have literally translated the Arabic expression mentioned in the first verse of the previous example. However, the result in English is not acceptable. For example, Yusuf Ali translated it as 'eat up the property', while Pickthall translated it as 'devour the wealth', as shown in Figure 2.13. The meaning of the word 'eating' in the context of the verses refers to a person who takes something that belongs to another (i.e. it is not their own). In Arabic, it does not literally mean the act of eating. The correct translation of this expression into English is 'Those who unjustly take possession of the orphans' properties...'.

**The Holy Quran [Sura: An-Nisa (4): Verse 10]:**

إِنَّ الَّذِينَ يَأْكُلُونَ أَمْوَالَ الْيَتَامَىٰ ظُلْمًا إِنَّمَا يَأْكُلُونَ فِي بُطُونِهِمْ نَارًا ۖ وَسَيَصْلَوْنَ سَعِيرًا ﴿10﴾

**The Holy Quran [Sura: At-Tawba (9): Verse 34]:**

... إِنَّ كَثِيرًا مِنَ الْأَحْبَارِ وَالرُّهْبَانِ لَيَأْكُلُونَ أَمْوَالَ النَّاسِ بِالْبَاطِلِ ... ﴿34﴾

**The Holy Quran [Sura: Al-Baqara (2): Verse 188]:**

وَلَا تَأْكُلُوا أَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ وَتُدْلُوا بِهَا إِلَى الْحُكَّامِ... ﴿188﴾

Figure 2.12: Examples of verses in which the words money and eat are mentioned next to each other.

The following points explain some features of the Quranic text (Abdul-Raof, 2013):

1. **Orthography**

**The Holy Quran [Sura: An-Nisa (4): Verse 10]:**

... إِنَّ الَّذِينَ يَأْكُلُونَ أَمْوَالَ الْيَتَامَىٰ ظُلْمًا ... ﴿10﴾

**Translator:** Yusuf Ali
**English Translation**: Those who unjustly eat up the property of orphans...(10)

**Translator:** Pickthalls
**English Translation**: Lo! Those who devour the wealth of orphans wrongfully... (10)

**The correct translation**
**English Translation**: Those who unjustly take possession of the orphans' properties...(10)

Figure 2.13: Example of Synthetic-Semantic Ambiguity.

Quranic orthography follows its own conventions and rules that differ from other Arabic script variants, such as DA, CA and MSA. The first rule is that many words do not adhere to known and agreed-upon Arabic spelling rules. Spelling differences in a word can occur by adding or deleting a letter, whether it is silent or spoken. These differences open up new meanings for words that are compatible with the context of the text. For example, in standard Arabic, the word زَكَاةً appears differently in the Quran, where it is correctly rendered as ٱلزَّكَوٰةَ. In standard Arabic, زَكَاةً means purity, while in the Quran, ٱلزَّكَوٰةَ means giving a portion of money to the poor according to certain conditions (Najjar, 2020). Second, there are Quranic words that are written the same way in Arabic but have different meanings in the Quran. For example, the word عَسَىٰ, which means wishing for something to happen, indicates a certainty of occurrence because it is from God in the Quran [1] [2] (bin Jarir al Tabari, 2001), as shown in Figure 2.14.

2. **Coherence**

[1] https://www.almaany.com/ar/dict/ar-ar/%D8%A7%D9%84%D8%B2%D9%83%D8%A7%D8%A9/
[2] https://kalemtayeb.com/quran/item/76855

**The Holy Quran [Sura: Al-Baqara (2): Verse 43]:**

وَأَقِيمُوا۟ ٱلصَّلَوٰةَ وَءَاتُوا۟ ٱلزَّكَوٰةَ وَٱرْكَعُوا۟ مَعَ ٱلرَّٰكِعِينَ ﴿43﴾

**Translator:** Yusuf Ali
**English Translation**: And be steadfast in prayer; practise regular charity; and bow down your heads with those who bow down (in worship (43).

**The Holy Quran [Sura: At-Tawba (9): Verse 102]:**

... عَسَى اللَّهُ أَنْ يَتُوبَ عَلَيْهِمْ ۚ ... ﴿102﴾

**Translator:** Yusuf Ali
**English Translation**: ... Perhaps Allah will turn unto them (in Mercy) ...(102).

Figure 2.14: Examples of feature Orthography challenges.

One of the challenges in understanding Quranic text is the coherence of its verses. To comprehend a verse or find an answer to a question posed about it, you may need to read a set of connected verses or an entire surah. For example, in the first verse of surah Al-Anfal, the Muslims asked the Prophet, may God bless him and grant him peace, how the spoils of war from the Battle of Badr would be divided. Subsequently, a series of verses (from verse 2 to verse 40) address various points, issues and reasons that Muslims must understand before they receive the answer. Specifically, the text discusses what characterises a true believer, the role of God and his angels in winning the battle and the believers' hesitation to fight. The answer to their question finally appears in verse 41, as shown in Figure 2.15 (Altammami, 2023b; Haleem, 2020).

## 2.4 The Hadith

During the mission of the Prophet Muhammad, may God bless him and grant him peace, which lasted 23 years, the Companions (Sahaba) memorised the Prophet's sayings and actions from the situations they experienced, but they did not write

**The Holy Quran [Sura: Al-Anfal (8): Verse 1-3]:**

يَسۡـَٔلُونَكَ عَنِ ٱلۡأَنفَالِۖ قُلِ ٱلۡأَنفَالُ لِلَّهِ وَٱلرَّسُولِۖ فَٱتَّقُواْ ٱللَّهَ وَأَصۡلِحُواْ ذَاتَ بَيۡنِكُمۡۖ وَأَطِيعُواْ ٱللَّهَ
وَرَسُولَهُۥٓ إِن كُنتُم مُّؤۡمِنِينَ ﴿1﴾ إِنَّمَا ٱلۡمُؤۡمِنُونَ ٱلَّذِينَ إِذَا ذُكِرَ ٱللَّهُ وَجِلَتۡ قُلُوبُهُمۡ وَإِذَا تُلِيَتۡ عَلَيۡهِمۡ
ءَايَٰتُهُۥ زَادَتۡهُمۡ إِيمَٰنٗا وَعَلَىٰ رَبِّهِمۡ يَتَوَكَّلُونَ ﴿2﴾ ٱلَّذِينَ يُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقۡنَٰهُمۡ يُنفِقُونَ ﴿3﴾

**Translator:** Pickthalls
**English Translation**: They ask thee (O Muhammad) of the spoils of war. Say: The spoils of war belong to Allah and the messenger, so keep your duty to Allah, and adjust the matter of your difference, and obey Allah and His messenger, if ye are (true) believers. (1) They only are the (true) believers whose hearts feel fear when Allah is mentioned, and when His revelations are recited unto them they increase their faith, and who trust in their Lord; (2) Who establish worship and spend of that We have bestowed on them. (3)

**The Holy Quran [Sura: Al-Anfal (8): Verse 41]:**

وَٱعۡلَمُوٓاْ أَنَّمَا غَنِمۡتُم مِّن شَيۡءٖ فَأَنَّ لِلَّهِ خُمُسَهُۥ وَلِلرَّسُولِ وَلِذِي ٱلۡقُرۡبَىٰ وَٱلۡيَتَٰمَىٰ وَٱلۡمَسَٰكِينِ وَٱبۡنِ
ٱلسَّبِيلِ إِن كُنتُمۡ ءَامَنتُم بِٱللَّهِ وَمَآ أَنزَلۡنَا عَلَىٰ عَبۡدِنَا يَوۡمَ ٱلۡفُرۡقَانِ يَوۡمَ ٱلۡتَقَى ٱلۡجَمۡعَانِۗ وَٱللَّهُ عَلَىٰ كُلِّ
شَيۡءٖ قَدِيرٌ ﴿41﴾

**Translator:** Pickthalls
**English Translation**: And know that whatever ye take as spoils of war, lo! a fifth thereof is for Allah, and for the messenger and for the kinsman (who hath need) and orphans and the needy and the wayfarer, if ye believe in Allah and that which We revealed unto Our slave on the Day of Discrimination, the day when the two armies met. And Allah is Able to do all things (41).

Figure 2.15: Example of Coherence from Quran.

them down in books. Instead, they passed them on verbally from generation to generation. Given the importance of the Prophetic Sunnah, as God commanded Muslims in the Holy Quran in many verses to follow the Sunnah of the Prophet Muhammad, may God bless him and grant him peace, including what is mentioned in Figure 2.16, Muslim scholars decided to record all hadiths in books for preservation. Since hadiths were not recorded directly during the lifetime of the Prophet, may God bless him and grant him peace, but were documented later, this led to the emergence of the science of Hadith. Hadith science studies the narrators, verifies them, and then confirms or rejects the hadiths. The importance of Hadith science lies in excluding false hadiths that have been fabricated about the Prophet, may God bless him and grant him peace.

The Hadith is considered the second primary source for Muslims in their religion. The term ‘hadith’ refers to the documentation of all aspects of the life of the Prophet Muhammad, may God bless him and grant him peace, including details such as words, actions, situations, events, conversations and mentions of previous stories. The lengths of hadiths vary, with some being short sentences and others being long passages.

**The Holy Quran [Sura: An-Nisa (4): Verse 59]:**

يَا أَيُّهَا الَّذِينَ آمَنُوا أَطِيعُوا اللَّهَ وَأَطِيعُوا الرَّسُولَ ... ﴿59﴾

**Translator:** Pickthalls
**English Translation**: O ye who believe! Obey Allah, and obey the messenger ...(59).

Figure 2.16: Example of Coherence from Quran.

### 2.4.1 Structure of the Hadith

One hadith consists of several sentences written in CA and is divided into two parts: the isnad and matn , as shown in Figure 2.17[1]. The isnad is metadata that describes the chain of narrators who transmitted the hadith. This part of

[1]https://sunnah.com/bukhari:1

the hadith is used solely to confirm its authenticity by verifying the reliability of the companions who narrated it. Religious scholars have focused on this aspect to classify the hadith as authentic or unauthentic. The matn is the core of the hadith and contains the sayings and deeds of the Prophet (Brown, 2017).

**Hadith Sharif:**
حَدَّثَنَا الْحُمَيْدِيُّ عَبْدُ اللَّهِ بْنُ الزُّبَيْرِ ، قَالَ : حَدَّثَنَا سُفْيَانُ ، قَالَ : حَدَّثَنَا يَحْيَى بْنُ سَعِيدٍ الْأَنْصَارِيُّ ، قَالَ : أَخْبَرَنِي مُحَمَّدُ بْنُ إِبْرَاهِيمَ التَّيْمِيُّ ، أَنَّهُ سَمِعَ عَلْقَمَةَ بْنَ وَقَّاصٍ اللَّيْثِيَّ ، يَقُولُ : سَمِعْتُ عُمَرَ بْنَ الْخَطَّابِ رَضِيَ اللَّهُ عَنْهُ عَلَى الْمِنْبَرِ، قَالَ : سَمِعْتُ رَسُولَ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ، يَقُولُ : " إِنَّمَا الْأَعْمَالُ بِالنِّيَّاتِ ، وَإِنَّمَا لِكُلِّ امْرِئٍ مَا نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى دُنْيَا يُصِيبُهَا أَوْ إِلَى امْرَأَةٍ يَنْكِحُهَا، فَهِجْرَتُهُ إِلَى مَا هَاجَرَ إِلَيْهِ "

**English Translation**:
Al-Humaydee `Abdullaah ibn Az-Zubayr narrated to us saying: Sufyaan narrated to us, who said: Yahyaa ibn Sa'eed Al-Ansaree narrated to us: Muhammad Ibn Ibraaheem At-Taymee informed me: That he heard 'Alqamah Ibn Waqaas Al-Laythee saying: I heard 'Umar ibn Al-Khattaab whilst he was upon the pulpit saying: I heard Allaah's Messenger (salallaahu `alaihi wassallam) saying: "The reward of deeds depends upon the intentions and every person will get the reward according to what he has intended. So whoever emigrated for worldly benefits or for a woman to marry, his emigration was for what he emigrated for".

Figure 2.17: Hadith example, Isnad in purple and Matn in black.

### 2.4.2 Al-Sihah Al-Sitta (the Six Books)

The religious scholar Muhammad Al-Bukhari collected hadiths and classified them as authentic or incorrect, then compiled them into a collection of books. He divided each book into chapters, and within each chapter, he grouped hadiths that discussed the same topic, naming the chapter accordingly. Some books are organised based on a series of narrators rather than topics. This collection of books is called the Sahih Al-Bukhari. The word 'sahih' means reliable and correct in Arabic. Many religious scholars following Sheikh Al-Bukhari also compiled hadiths into books. The six most important books specialising in hadiths are called the Al-Sihah Al-Sitta (in English, the Six Books): Sahih Al-Bukhari[1], Sunan

[1]Mohammad Al-Bukhari, Sahih Al-Bukhari (2002). Damascus: Dar Ibn Kathir.

| **Book** | **# of Hadiths** |
|---|---|
| Muslim | 7,293 |
| Abu Daud | 5,141 |
| Tirmizi | 4,209 |
| Ibn Maja | 10,082 |
| Nesa'i | 5,680 |
| Bukhari | 6,633 |
| **Total** | 39,038 |

Table 2.1: The number of hadiths in each book from Al-Sihah Al-Sitta books.

Abu Dawood[1], Sunan Al-Nasai[2], Sunan Ibn Maja[3], Sahih Muslim[4], and Sunan Al-Tarmithi[5]. The number of hadiths in each book is listed in Table 2.1 (Al-tammami *et al.*, 2020; Brown, 2017). The Hadith is available on several websites, including sunnah.com, dorar.net, and hadithcollection.com.

### 2.4.3 Challenges in the Hadith

The following points explain some features of the Hadith :

1. **Metaphor**

   A metaphor involves carrying the meaning of one word over to another. Runcie *et al.* (2000) define metaphor as 'a way of describing something by comparing it to something else which has the same qualities, but without using the words "as" or "like"'. An example of a metaphor is shown in Figure 2.18. In this example, women are likened to glass vessels because they are considered sensitive, weak and fragile (Athman *et al.*, 2015; ZELACI, 2014).

2. **Simile**

[1] Abu Dawood, Sunan Abu Dawood (2017). Damascus: Dar Ibn Kathir.
[2] Imam Al-Nasai, Sunan Al-Nasai (2017). Damascus: Dar Ibn Kathir.
[3] Ibn Majah, Sunan Ibn Majah (2016). Damascus: Dar Ibn Kathir.
[4] Muslim Ibn Al-Hajjaj, Sahih Muslim (2017). Damascus: Dar Ibn Kathir.
[5] Al-Tarmithi, Sunan Al-Tarmithi (2016). Damascus: Dar Ibn Kathir.

**Hadith Sharif:**
حَدَّثَنَا أَبُو الرَّبِيعِ الْعَتَكِيُّ، وَحَامِدُ بْنُ عُمَرَ، وَقُتَيْبَةُ بْنُ سَعِيدٍ، وَأَبُو كَامِلٍ جَمِيعًا عَنْ حَمَّادِ بْنِ
زَيْدٍ، قَالَ أَبُو الرَّبِيعِ حَدَّثَنَا حَمَّادٌ، حَدَّثَنَا أَيُّوبُ، عَنْ أَبِي قِلاَبَةَ، عَنْ أَنَسٍ، قَالَ كَانَ رَسُولُ اللَّهِ
صلى الله عليه وسلم فِي بَعْضِ أَسْفَارِهِ وَغُلاَمٌ أَسْوَدُ يُقَالُ لَهُ أَنْجَشَةُ يَحْدُو فَقَالَ لَهُ رَسُولُ اللَّهِ
صلى الله عليه وسلم " يَا أَنْجَشَةُ رُوَيْدَكَ سَوْقًا بِالْقَوَارِيرِ " .

**English Translation**:
Anas reported that Allah's Messenger (ﷺ) had in one of his journeys his black slave who was called Anjasha along with him. He goaded by singing the songs of camel-driver. Thereupon Allah's Messenger (ﷺ) said:

Anjasha, drive slowly as you are driving (the mounts who are carrying) glass vessels.

Figure 2.18: Example of metaphor.

A simile is explicit and clear, using words such as 'like', in contrast to a metaphor, which is implicit. In the hadith in Figure 2.19, the Prophet Muhammad, may God bless him and grant him peace, describes the revelation as having the sound of a bell (Dina, 2008).

## 2.5 The Arabic Language

Arabic is considered the fourth most used language on the Internet. It is spoken by over 400 million people (Guellil *et al.*, 2021), and nearly 1.4 billion Muslims around the world use it in their prayers. Arabic consists of 28 letters. The structure and nature of the Arabic language are characterised by several features. Unlike many other languages, the direction of writing in Arabic is right to left. There are no capital letters to distinguish proper names, which makes it more difficult to recognise them in text. For example, in English, the letter 'A' in capitalised form and the letter 'a' in lowercase represent the same letter; the name 'Adam' is written with an initial capital letter because it is a name. In Arabic, the first letter, Alif, is written in one form, 'أ', so Adam's name is written as 'أدم'. In addition, diacritics are used in Arabic to help determine the correct meaning of a word, and the shape of a letter changes depending on its position in

**Hadith Sharif:**

وَحَدَّثَنِي عَنْ مَالِكٍ، عَنْ هِشَامِ بْنِ عُرْوَةَ، عَنْ أَبِيهِ، عَنْ عَائِشَةَ، زَوْجِ النَّبِيِّ صلى الله عليه وسلم
أَنَّ الْحَارِثَ بْنَ هِشَامٍ سَأَلَ رَسُولَ اللَّهِ كَيْفَ يَأْتِيكَ الْوَحْىُ فَقَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم
" أَحْيَانًا يَأْتِينِي فِي مِثْلِ صَلْصَلَةِ الْجَرَسِ وَهُوَ أَشَدُّهُ عَلَىَّ فَيَفْصِمُ عَنِّي وَقَدْ وَعَيْتُ مَا قَالَ وَأَحْيَانًا
يَتَمَثَّلُ لِيَ الْمَلَكُ رَجُلاً فَيُكَلِّمُنِي فَأَعِي مَا يَقُولُ " . قَالَتْ عَائِشَةُ وَلَقَدْ رَأَيْتُهُ يَنْزِلُ عَلَيْهِ فِي الْيَوْمِ
الشَّدِيدِ الْبَرْدِ فَيَفْصِمُ عَنْهُ وَإِنَّ جَبِينَهُ لَيَتَفَصَّدُ عَرَقًا .

**English Translation**:

Yahya related to me from Malik from Hisham ibn Urwa from his father from A'isha, the wife of the Prophet, may Allah bless him and grant him peace, that al-Harith ibn Hisham asked the Messenger of Allah, may Allah bless him and grant him peace, "How does the revelation come to you?" and the Messenger of Allah, may Allah bless him and grant him peace, said, "Sometimes it comes to me like the ringing of a bell, and that is the hardest for me, and when it leaves me I remember what it has said. And sometimes the angel appears to me in the likeness of a man and talks to me and I remember what he says".

A'isha added, "I saw it coming down on him on an intensely cold day, and when it had left him his forehead was dripping with sweat."

Figure 2.19: Example of simile.

a word. For example, if the letter 'H' is at the beginning of a word, it is written 'هـ'; in the middle of a word, it is written 'ـهـ'; and at the end of a word, it is written 'ـه' (Farghaly & Shaalan, 2009).

The Arabic language can be divided into three major forms: CA, MSA and DA. Figure 2.20 shows the relationship between these types. An Arab person usually uses the three types daily. He/she uses CA language in prayer and when reading religious and historical books. He/she uses dialects in his/her daily life at home and uses the MSA language to read the news, view television, participate in courses, and be part of a work environment (Shaalan *et al.*, 2019).

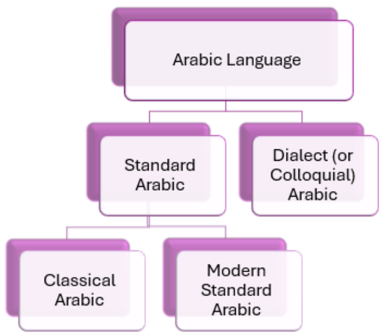



Figure 2.20: Categories of Arabic language.

### 2.5.1 Classical Arabic

CA was the language used among Arabs at the time of the Prophet Muhammad, may God bless him and grant him peace; consequently, the Quran was revealed and the Hadith was written in it. Today, CA is considered the language of the elite, as well as the language of science and religious and historical books. Learning CA is important for reading and understanding the Holy Quran, the Hadith, and history books (Huehnergard *et al.*, 2013).

### 2.5.2 Modern Standard Arabic

MSA is currently the official language of 21 countries. It is derived from CA and differs in terminology and vocabulary but follows the same linguistic rules. An example of MSA is shown in Figure 2.21[1]. MSA is used widely in university teaching, seminars, learning courses, books, media, newspapers and in workplaces (Abdelali, 2004; Alqadasi *et al.*, 2023; Altammami, 2023b).

| id | title | context | question | answers |
|---|---|---|---|---|
| 130251388427 | السعودية | السعودية ورسميًا المملكة العربية السعودية هي أكبر دولة في الشرق الأوسط وتقع تحديدًا في الجنوب الغربي... | ما هي الحدود في - الشمال؟ ال | {'text': array(['يحدها من الشمال العراق والأردن'], dtype=object), 'answer_start': array([199], dtype... |
| 791276760553 | السعودية | السعودية ورسميًا المملكة العربية السعودية هي أكبر دولة في الشرق الأوسط وتقع تحديدًا في الجنوب الغربي... | ما هي الحدود في - الشمال الشرقي؟ ال | {'text': array(['الكويت'], dtype=object), 'answer_start': array([237], dtype=int32)} |

Figure 2.21: An example of MSA text.

### 2.5.3 Dialectal Arabic

DA consists of the informal language varieties used for conversations in Arab countries stretching from the Persian Gulf in the east to the Atlantic Ocean in the west (Elnagar *et al.*, 2021). The dialects of these countries can be divided into five main categories: Levantine, Egyptian, Moroccan, Iraqi and Gulf (Habash, 2010). For example, the word 'sofa' in English corresponds to different vocabulary in various Arabic dialects: Levantine 'كنبايه', Egyptian 'كنبه', Moroccan 'فوطوي', Iraqi 'كرويت', and Gulf 'باطرمه' (Aliwy *et al.*, 2020).

[1] https://www.kaggle.com/datasets/thedevastator/unlocking-arabic-language-comprehension-with-the

## 2.6 Conclusion

This chapter reviewed the religious background that the reader must understand in relation to this research regarding Islamic books, the features of their texts and the challenges that AI models face when processing them. In addition, it describes the Arabic language and its main types.

# Chapter 3

# Literature Review

## 3.1 Introduction

This chapter, based on my publication (Alnefaie *et al.*, 2022b), meticulously examines the progression and methodologies employed in NLP to construct a QA system grounded in the Holy Quran and Hadith. Additionally, it provides a critical evaluation of the deficiencies and limitations within existing systems.

## 3.2 Religious Question Answering Systems

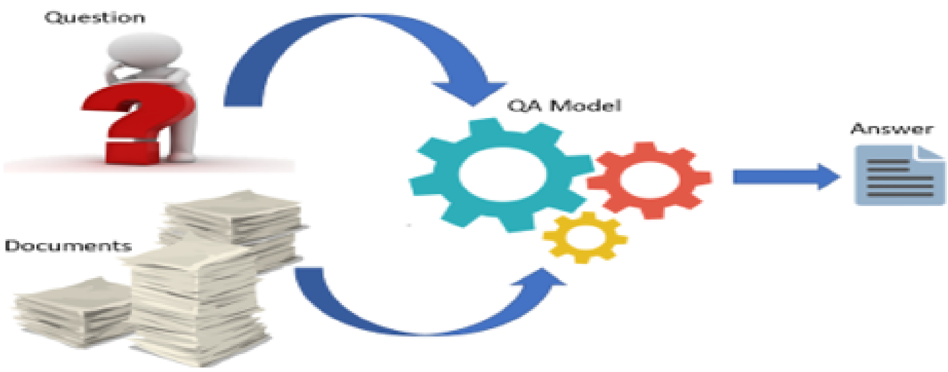



Figure 3.1: The question and answer system structure.

Numerous systems have been developed to extract knowledge from holy books, as illustrated in Figure 3.1. These systems can be categorised into two types, as depicted in Figure 3.2. The first type concentrates on retrieving information from the texts of the Quran or Hadith, which can be referred to as retrieval-based systems. This type relies completely on NLP tools and techniques to find answers. In contrast, knowledge-based systems focus on creating a dataset that encompasses the largest possible number of questions and their corresponding answers from the Quran and Hadith and then utilise either retrieval techniques or a pre-trained DL model to provide answers to users.

The retrieval techniques used in these systems can be divided into text-based and semantic-based methods, as shown in Figure 3.3 (Alqahtani, 2019). Text-based techniques retrieve information by matching the words in a provided question, while semantic-based techniques return answers by matching the concept or meaning of a user's query (Sudeepthi *et al.*, 2012). Most applications and websites employing Quran and Hadith search tools use text-based techniques. Keyword matching (Munshi *et al.*, 2022) and chatbots (Sihotang *et al.*, 2020) are two commonly used techniques, although there are many others. The keyword-matching approach uses the terms in a question directly during the search process, while the chatbot technique selects the most relevant terms from a question and uses those terms in the search process.

Semantic-based techniques can be divided into synonym sets, ontology-based methods and cross-language approaches (Alqahtani, 2019). Each of these techniques employs a different method to search for meaning. For example, the synonym technique extracts all synonyms of the query terms from external sources, such as WordNet, and then adds these synonyms to the query (Neamah & Saad, 2017). An ontology-based technique identifies a concept that matches the user's question and retrieves the relevant text associated with this concept (Khan *et al.*, 2013). The idea behind cross-language techniques is to translate the query into other languages and use these new queries to retrieve answers by applying a matching process (Yunus *et al.*, 2010).

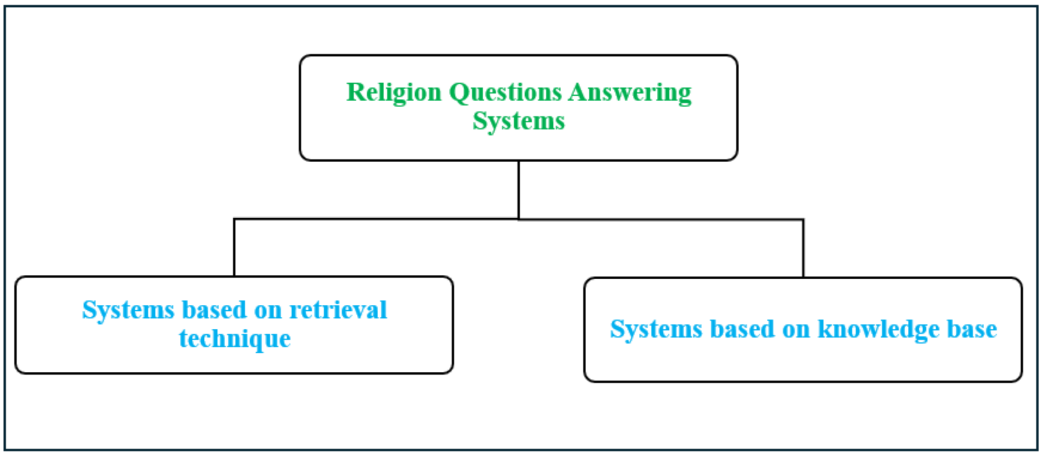



Figure 3.2: The classification of the religion questions answering systems.

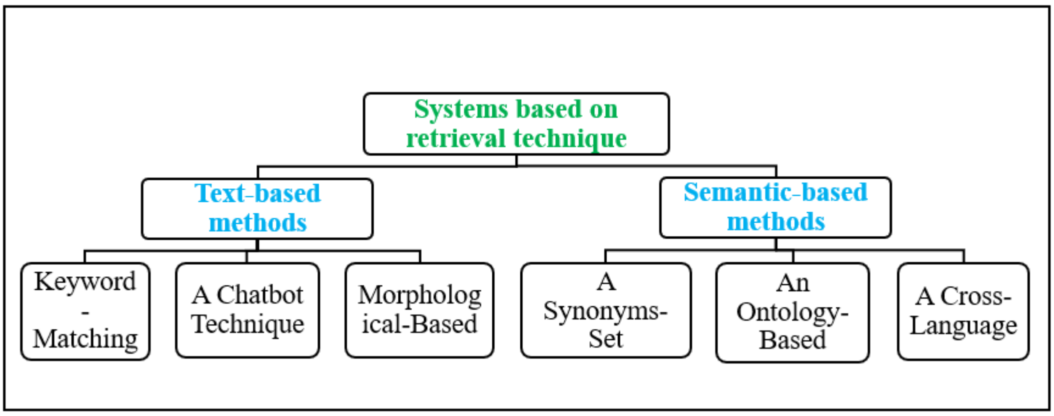



Figure 3.3: The classification of the systems based on retrieval techniques.

## 3.3 Religious Web Application Search Tools

Many websites aim to extract answers from the Hadith or the Quran. The text-based method is used in most of these websites, as described below.

Several Quran websites are available online, such as the Noble Quran[1], Quranic Arabic Corpus (QAC)[2], Tanzil[3], KSU Quran[4], and The Quran[5]. These websites allow users to explore the Quran in various formats, such as text and audio, in several languages.

Moreover, these tools enable users to search using multiple options, such as by words, roots, chapter number, surah number and verse number, making it easier to retrieve and access the required information. The search options for words and roots allow users to retrieve verses containing the exact query words or any words with the same root. For instance, the verses containing 'يدعون', 'دعاء' and 'دعوة' will be returned when the query is 'دعاء', as shown in Figure 3.4.

The Quranic Arabic Corpus [6] is a website that allows users to search the Holy Quran by word or concept. An ontology was built as part of this site in which verses related to each concept were collected and indexed under them, as shown in Figure 3.5.

Many websites specialise in retrieving knowledge from the hadiths of the Prophet, such as Search Truth Tool[7], Hadeethenc[8], Dorar[9], and Sunnah[10].

The Search Truth Tool allows users to search using English words in the following hadith books: Malik's Muwatta, Sahih Muslim, Sahih Al-Bukhari and Sunan Abu-Dawud. This tool enables searches using more than two words. Users can search in two ways: by the exact words of the query or by the root of the query words.

[1] http://quran.com/
[2] http://corpus.quran.com/
[3] http://tanzil.net/
[4] http://quran.ksu.edu.sa/
[5] http://thequran.info/
[6] https://corpus.quran.com/
[7] https://www.searchtruth.com/
[8] https://hadeethenc.com/ar/home
[9] https://www.dorar.net/
[10] https://sunnah.com/

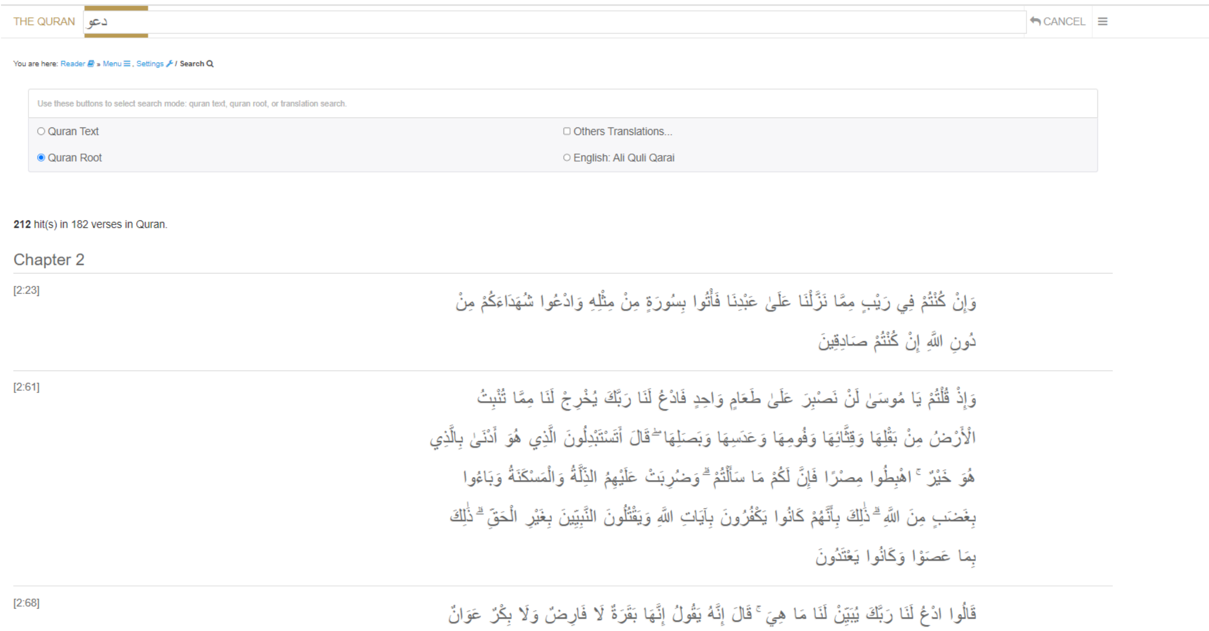



Figure 3.4: An example of searching by roots using the The Quran website.

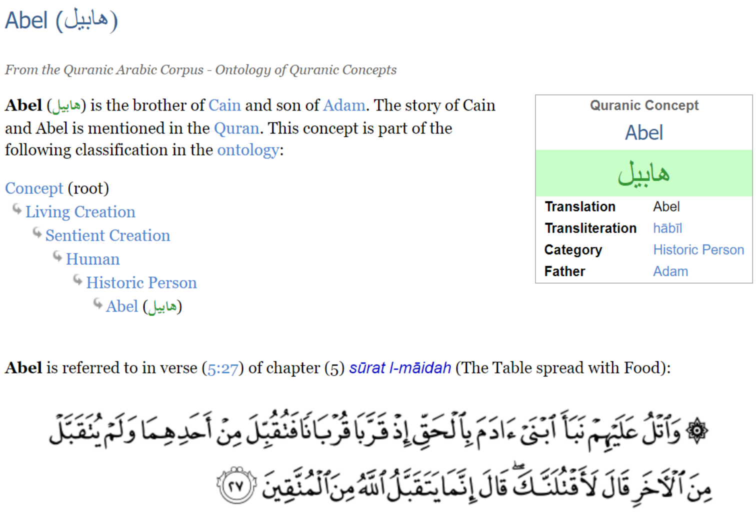



Figure 3.5: An example of searching by concepts using the Quranic Arabic Corpus website.

Hadeethenc is an integrated project that selects repetitive hadiths and their explanations and then provides a high-quality translation of them in many languages. It also provides a retrieval tool that allows users to search using words in several languages.

Dorar has built a comprehensive electronic database for the inheritance of the Messenger Muhammad, peace be upon him, and has facilitated access to it through a search engine that allows users to enter words in both Arabic and English to retrieve relevant information.

Sunnah is a website that supports the search for both English and Arabic Hadith from many hadith books, such as Sahih Muslim, Musnad Ahmad, Muwatta Malik, Sunan Ibn Majah, Jami at-Tirmidhi, Sunan Abi Dawud and Sunan an-Nasa'i.

Nevertheless, these web applications fail to retrieve verses or hadiths that contain words that are different from the query but have the same meaning. The Quranic Arabic Corpus website, which provides a concept search feature, is limited to concepts covered by the ontology. In addition, it is a simple word search tool based on text-based techniques. It cannot find answers from verses or hadiths for different types of questions. For example, it cannot answer the question, "What is permissible and forbidden?" 'ماهو الحلال والحرام', as shown in Figure 3.6. When you type the word 'permissible' (halal) or 'forbidden' (haram) alone in the search box, the site retrieves texts containing either 'halal' or 'haram.' However, when you enter the question 'What is permissible (halal) and forbidden (haram)?'—which requires a comprehensive and detailed answer about halal and haram actions in Islam—the site fails to provide an answer and returns zero results.

## 3.4 Religious Question Answering Systems Research

A significant number of QA studies have been conducted in the domain of the Quran and Hadith, producing systems based on retrieval techniques or knowledge-based systems, as outlined below.


Figure 3.6: the result of the question over the Dorar search tool.

### 3.4.1 Systems Based on Retrieval Techniques

1. **Text-Based Search Technique:**

   Shmeisani *et al.* (2014) built a system to answer questions about the Quran in the Indonesian language . It retrieved relevant documents using a keyword-based approach and extracted answers using a rule-based method. Nevertheless, this system answered only a specific type of question related to the Al-Baqarah Surah.

   AbuShawar & Atwell (2016) proposed a Quran chatbot system, which allows users to type and receive answers to English questions using both Arabic and English Quran verses. ALICE (artificial linguistic Internet computer entity) was the chatbot platform used to implement this system, which employed pattern-matching techniques.

   Hassan *et al.* (2017) recommended the Al-Hadith search engine tool in four different languages: Russian, French, Arabic and English. It was built in two stages: corpus construction and retrieval model building. In the corpus construction stage, the Multilingual Hadith Corpus (MHC) was created from the Sahih Al-Bukhari. A total of 2,030 Arabic hadiths were identified along with their translations into the other three languages. After the data were cleaned and preprocessed, language specialists reviewed the hadiths.

Based on cosine similarity measures, a retrieval module was developed to retrieve relevant hadiths related to the query. A website for this tool was created using Microsoft Visual Studio 2012. Nevertheless, the collection contained only 2,030 hadiths, while the total number of hadiths was approximately 14,000.

Adany *et al.* (2017) developed several prototypes to retrieve verses that answer user questions from the chapters of Al-Baqarah and Al-Fatihah. The first prototype is a basic NLP model consisting of tokenising a query, comparing the tokenised terms with the Quran text and displaying the matched verses. The experimental results of this prototype showed that unnecessary verses were included in response to queries. He proposed solving the problem by removing some diacritics, stop words and punctuation. The results indicated that performance improved after making these adjustments.

There is a type of plural in the Arabic language where a specific pattern is used when pluralizing a word. For example, when pluralizing the word exegesis تفسير or exercise تمرين, we use the pattern تفاعيل. This means that to write these words in the plural form, we need to add the letter alif (ا) to the third letter of the word. The plural form of exegesis تفسير, after adding the letter alif, becomes exegeses تفاسير, and the plural form of exercise تمرين becomes exercises تمارين.

Sometimes, if the question contains a word in the plural form and the Qur'an contains the same word in the singular form, the system fails to retrieve the verse. For example, if the question includes the word exegeses تفاسير, the system will fail to retrieve the verse: 'And they do not come to you with an argument except that We bring you the truth and the best explanation' وَلَا يَأْتُونَكَ بِمَثَلٍ إِلَّا جِئْنَاكَ بِالْحَقِّ وَأَحْسَنَ تَفْسِيرًا, because it contains the same word but in the singular form.

The second prototype consisted of three steps. First, it analysed the question by stemming and removing stop words. It then generated more keywords using the تفاعيل pattern. In other words, when the question contained a term with a weight of تفاعيل, this singular term was added to the question. Next, the search process was applied to find answers from the Quran verses stored using Lucene indexing. The third prototype aimed to increase precision by using three patterns: تفاعيل,فياعيل and فعاعيل (instead of just one pattern as in the second prototype). The fourth prototype utilised the exaggeration formula pattern, while the last prototype applied plural, dual and singular forms. The results showed that recall and precision were enhanced with the addition of all these patterns.

Sihotang *et al.* (2020) created an Indonesian chatbot system to answer Islamic questions utilising the fuzzy string-matching algorithm. The system dataset contained question–answer data about the topics of zakat and prayer collected from websites.

2. **Semantic Search Technique:**

   - **An Ontology-Based Technique:**

     Many studies have sought to find meaning by identifying the concept corresponding to the words in a question in an ontology and then retrieving all verses associated with that concept. An ontology in this context is defined as 'an explicit specification of a conceptualisation of a domain, in terms of concepts, attributes and relations' (Alqahtani, 2019). Entities in a given domain are represented by concepts in the ontology, which are organised in a taxonomy tree. In this tree, the concepts are linked by relationships, as shown in Figure 3.7 [1].

     Abbas (2009) developed a search tool for the Quran in Arabic and English called Qurany. The Quran texts were available in this tool in Arabic, with eight parallel English translations. It featured two modules: keyword search and abstract concepts. The keyword search tool allowed queries in Arabic as well as queries in English. This tool expands English queries using WordNet synonyms (synonyms in Arabic could not be used due to the limitations of Arabic in WordNet at that time). Additionally, it supported searching for lemmas and

[1] https://corpus.quran.com/concept.jsp

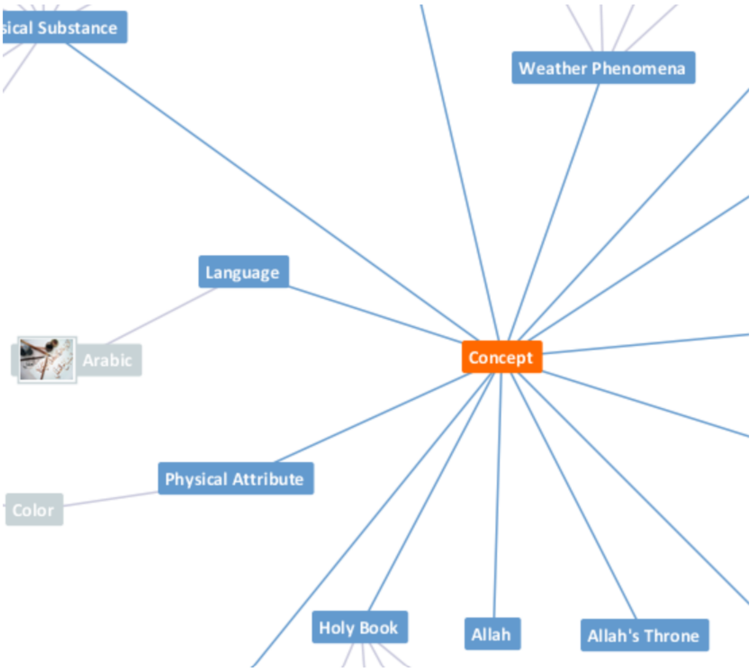



Figure 3.7: The Quranic Ontology.

morphemes instead of only keywords. The abstract concepts module allowed users to search by concept using the Qurany ontology. The Qurany ontology was a tree of all Quran concepts, structured in several levels: the first level contained the main concepts, the next level contained the sub-concepts, and so on, until the leaves were reached. The leaves contained verses related to the concepts above them in the tree. The Qurany ontology was built based on Mushaf Al Tajweed's book (Habash, 2001). Unfortunately, this tool is currently unavailable.

Khan *et al.* (2013) developed a Quran semantic search using an ontology built for animals and birds, which appeared in the book Al-Hayawany Fi el Quran Al-Kareem (El-Nagga, 2006). There are around 167 animals mentioned in the Quran. These researchers proposed building an Islamic and Quranic WordNet because many Islamic concepts are either not mentioned in the Arabic WordNet or do not have the correct meaning, as they have been defined by non-Muslims. As a result, search performance improved by expanding queries using ontology.

Shmeisani *et al.* (2014) recommended a semantic model based on ontology to find answers to user questions from the Arabic Quran. The system consists of three layers: question processing, semantics and query building. A question was pre-processed in the question processing layer by removing special characters and stop words, determining the part of speech (POS) tags and identifying the answer type or domain, such as who, where or when. After that, the semantic layer attempted to enhance the system's performance by understanding the question's meaning, not just the syntax. In other words, if the question's words did not appear in the ontology, synonyms from the Al-Maany dictionary were added to the query. Finally, a SPARQL query was built and executed against the Quranic ontology to retrieve the answer.

The cited paper constructed the Quran ontology from the text using the Arabic ontology extractor approach. This approach consisted of three general steps: specifying the patterns used to extract the two concepts (two nouns) and the relation (a verb) between them, determining all the instances that matched these patterns in the Quran text and storing all the instances collected in the ontology. They used three patterns in the first step: two nouns before a verb, two nouns after a verb and a verb between two nouns. The basic idea of choosing these patterns was based on the hypothesis that a relation is often expressed with a verb and a concept with a noun. They found approximately 380 concepts, such as 'book' and 'messenger', and 50 relations, such as 'show' and 'live'.

However, the system only answered factoid questions involving a number, name, or date. These were simple questions, such as 'where,' 'when,' etc.

Abdelnasser *et al.* (2014) developed a Quran Arabic QA system called Al-Bayan. The input to this system was a user query, while the output was a Quran passage that included the answer and tafseer from the interpretation books. The system architecture was created in two phases: online and offline. Building the ontology and creating the inverted index occurred in the offline phase. This system's ontology was created by integrating the Qurany Ontology (Abbas, 2009) and the Quranic Corpus Ontology (Atwell *et al.*, 2011). It was a tree of Quranic concepts that showed the relationships between them. After the integration process, 621 out of 6,236 verses did not belong to any concept. They added these verses under the most related concept, using a similarity measure. The produced ontology had 217 leaf concepts. Under every concept, they stored the verses and their tafseers. In the process of creating the inverted index, the list of verses under each concept was pre-processed using the Morphological Analysis and Disambiguation for Arabic (MADA) toolkit (Habash *et al.*, 2009). For each word, MADA generated a stem and POS tag. It then created a vector of terms that appeared under each concept and weighted these vectors using the term frequency-inverse document frequency (TF-IDF) scheme for each concept. Finally, they linked each term to the list of concepts that appeared under it in the ontology by constructing an inverted index.

The online phase involved analysing the question, retrieving the information and extracting the answer. There were two modules in the question analysis step: question pre-processing and question classification. The question pre-processing step resulted in a vector of terms in which each term included a POS tag and MADA stem. They classified the questions to determine the expected answer type and reduce the search scope. They used a support vector machine (SVM) classifier to determine the question type, such as whether it involved searching for a person's name or a place. After that, the verses most relevant to the query and their tafseers were retrieved from the ontology. Answer extraction was the final step. They added named entity recognition (NER) tags to the relevant verses and their tafseers using the LingPipe tool[1]. The text was then split into passages, with each considered a candidate answer. The candidate answers were ranked based on features, such as the number of matched terms between candidate answers and questions and whether the named entity in the candidate answer fit the question type (e.g. 'where' as a question type would fit 'location' as a named entity of a candidate answer). Five Quran experts evaluated the system.

However, if the words used in the question were not found in the ontology concepts, the system would not return any answers. Moreover, it could only answer questions about locations, numbers, descriptions, physical entities or creation.

Hassan & Atwell (2016) proposed a concept search tool in the Al-Hadith area. Development of this system had three phases. The first phase involved defining the dataset. Sahih Bukhari's book was divided into 100 concepts, with several hadiths under each concept. The hadiths were organized into thematic sections, with each section representing a distinct concept in Islamic teachings. This categorization allows scholars to efficiently find hadiths that pertain to different areas of life, faith, and practice. The researchers selected a collection of hadiths; extracted their concepts; translated the hadiths into Russian, Arabic, French and English; and used this book structure to build the website. The second phase involved creating HTML files, where a page was allocated to each hadith. In the third phase, a website was constructed for this tool. The tool presented the results as a tree of concepts for each language. Users could browse the main and sub-concepts, and the leaf concept contained a list of hadith texts related to this concept. They used 10 concepts from these trees to evaluate their system. Zulkefli *et al.* (2015) also developed a similar idea for hadith in the Malay language. However, the concepts and hadith numbers were very limited in both studies, with around 100 concepts and, in some languages, fewer than 100.

[1] http://alias-i.com/lingpipe

Ahmad *et al.* (2017) recommended a Malay Quran system based on an ontology. The system stages included ontology creation and query processing. In the ontology creation stage, they translated English and Arabic ontologies, such as the Qurany ontology (Abbas, 2009) and Quranic ontology (Dukes, 2015), into Malay using a bilingual dictionary and WordNet. Experts in the Malay Quran then reviewed them. In the query processing stage, there was an NLP and semantic analysis phase. The NLP phase consisted of removing stop words, stemming the words and adding word lemmas. After the question analysis using the NLP phase, the semantic analysis technique was applied using ontologies to answer the questions.

Alqahtani (2019) and Alqahtani & Atwell (2015, 2016b, 2017) reviewed existing search systems in the Quran area and evaluated them using 13 criteria. They found many shortcomings, such as each system using only one ontology and the existing ontologies not covering all Quran topics, which affected search system performance. Therefore, Alqahtani (2019) and Alqahtani & Atwell (2016a, 2018) developed an ontology by aligning and integrating three existing ontologies: QurAna (Sharaf & Atwell, 2012), Qurany Ontology (Abbas, 2009) and the Quranic Ontology (Dukes, 2015). Additionally, they utilised nine different resources to enrich the ontology, such as Tafseer Al-Mussar and Tafseer Al-Jalalain. Finally, Alqahtani (2019) proposed using this new ontology as a basis for the Arabic Quran search model.

The Arabic Quran search model consisted of five components: an Arabic question analyser module, an Arabic similarity Quranic words model, a semantic search model, a keyword model and a scoring and ranking model. A user's question was input into the Arabic question analyser module, which went through several steps: determining the user question class using a classifier, stemming the question terms, adding POS tags to recognise the nouns and verbs in the question and generating synonyms for nouns and verbs based on the Arabic similarity Quranic words model. After that, the semantic search model retrieved the most relevant concepts to the question using the new ontology and sent these concepts' verses to the scoring and ranking model. Additionally, if no verse was retrieved, the keyword model retrieved any verse matching the question. Finally, the scoring and ranking model ranked the candidate verses based on the frequency of the question terms in the verse. According to Alqahtani (2019), despite the system's good performance, there were shortcomings in several parts. First, the ontology did not cover many aspects, such as laws, principles and Islamic ethics. Additionally, there was a weakness in the classifier's performance, which required an increase in the training dataset size. The Arabic semantic similarity Quranic word model required further research and improvement to generate words similar to those in the question.

- **A Synonym-Set Technique:**

  Many researchers have added synonyms of question words to a query to search for the broader meaning of a word rather than its actual use in a text. This technique is known as 'query expansion' (Carpineto & Romano, 2012). To find these synonyms, they used various approaches, including statistical methods and methods that utilised linguistic knowledge resources such as dictionaries and WordNet. WordNet is a lexical database concerned with words and their meanings, similar to a dictionary (Miller, 1995). What distinguishes WordNet is that it groups synonyms into sets called synsets, whereas a typical dictionary arranges words alphabetically (Banerjee & Pedersen, 2002).

  Neamah & Saad (2017) developed a dataset of hadiths from Al-Bukhari on specific topics and created a QA system using this dataset. When constructing the dataset, the researchers used a classification algorithm to divide the Arabic hadiths into two categories: prayer and fasting. Next, they used NER to classify the prayer and fasting categories into subcategories based on time and place topics. The result was four categories: when for fasting, where for fasting, where for prayer and when for prayer. This narrowed the research scope, which positively affected the system's performance. The system expanded the submitted questions using WordNet. Nevertheless, this system only answered where and when questions concerning prayer and fasting.

Yusuf *et al.* (2019) built a system to answer Quran questions using WordNet. The researchers used WordNet to extract synonyms for the question words and added these synonyms to a submitted question to form a new query, which they subsequently used in the system. They tested the model using three different English translations of the Quran: Arberry, Sarwar and Yusuf Ali. On average, the system's evaluation results performed better than three traditional models: best matching (BM25), TF-IDF and Lucene.

Yusuf *et al.* (2020) recommended a Quran semantic search tool in French to improve performance. They expanded a query by including the most co-occurring terms in the Quran along with the query terms. The performance of this system was better than that of the BM25 and Yusuf *et al.* (2019) models, achieving a 36% improvement.

Abdi *et al.* (2020) recommended building a system that answers questions about Al-Hadith by utilising an Arabic dictionary. The system calculated a score for each sentence in the Hadith text dataset, which contained around 4,000 hadiths from the Sahih al-Bukhari books. The equation for calculating this score included two parameters. The first parameter was the semantic similarity value between the sentence and the question, while the second parameter was the degree of semantic similarity between the sentence and other sentences related to the question. Semantic similarity was calculated using synonyms from an Arabic lexical database and the order of the words in the sentences. This system achieved good performance; however, the authors noted that the coverage of the lexical database was limited, which affected performance. Moreover, the system provided only partial answers to non-factoid questions.

Maraoui *et al.* (2021) proposed answering Arabic factoid questions with a system that operated in three stages: question processing, information extraction and answer processing. In the first stage, the question was analysed to add more detailed information that described the user's request, including the question topic, NER tag and similar questions. A dictionary was used to determine the topic of the question. Moreover, they added questions that were synonymous with a user's question in meaning but had a different syntax from a second dictionary using pattern matching techniques. A Hadith text dataset was constructed using the Text Encoding Initiative (TEI) standard, which contains text with tag descriptions. Therefore, answers were extracted from the dataset using these tags. Finally, the system formulated an answer. However, it only answered factoid questions.

- **A Cross-Language Technique:**
  Yunus *et al.* (2010) proposed a Quran QA system that supported Malay, Arabic and English. This system was based on retrieving more relevant documents by improving the query using different languages. First, the user entered a query. The system then translated the question into the other two languages. Finally, the relevant documents from the three language datasets were retrieved using pattern-matching techniques. However, the results showed that irrelevant documents were returned, and the researchers attempted to enhance performance by adding query synonyms from the dictionary after translation for each language.

Systems based on retrieval techniques have many shortcomings. A text-based method may fail to retrieve related documents containing synonyms for the query words because it relies on retrieving documents that contain only the actual query words. It does not consider the meaning of the query words during the retrieval process. In comparison, a semantic search method may struggle to provide the correct answer because it is based on incomplete external resources, such as ontologies and dictionaries. These resources do not cover all areas of a domain and do not encompass all concepts, terms and relationships. Therefore, some studies have sought to integrate more than one ontology to provide a broader semantic context.

One of the most significant weaknesses of existing systems is their ability to answer only factoid questions; none found in the literature was able to address other complex question types. A typical QA system takes a query from the user, passes it through several components and then extracts the answer from a dataset. The tasks of these components include pre-processing the question to analyse and attempt to understand it as comprehensively as possible, representing the question in a specific format that a machine can comprehend and then using various techniques to find the answer from the knowledge representation. There are weaknesses in existing natural language analysis tools and approaches that pose one of the most prominent current challenges. Consequently, even state-of-the-art systems face obstacles in answering all types of possible questions.

### 3.4.2 Systems Based on a Knowledge Base

Recently, attempts have been made to build a knowledge base to answer all question types. The knowledge base used is a collection of questions and answers. It can be classified as a closed-domain or open-domain knowledge base, depending on the scope of the content. A closed-domain knowledge base is built to answer questions in a specific domain. In contrast, an open-domain knowledge base attempts to answer questions in any area.

This section focuses on knowledge base systems, which we can further divide into systems that use traditional retrieval techniques or deep-learning models.

1. **Knowledge Base System using Traditional Retrieval Technique**

   Saeedi *et al.* (2014) introduced Quranjooy, a closed-domain knowledge base QA system for the Persian Quran. The researchers gathered approximately 6,000 Persian question–answer pairs from credible websites to develop the knowledge base for the system. As with typical QA system architectures, this system comprises three traditional stages. First, the system analysed a question to obtain detailed information, such as the question type. Second, four modules were applied in parallel to retrieve an answer from the knowledge base. Finally, the outputs of these modules were merged to obtain the correct answer. The modules in the second stage included NER, verse finder, tabular and ontology-based methods. Each module retrieved candidate answers and scored them based on their relevance to the question. The most relevant answer was then considered the output of the module. The NER module could answer nine types of questions. For example, if the question type was a date, this module focused the search process on words tagged as dates using NER. The second module could answer questions about verse references and the frequency of a specific term in the Quran. Questions regarding interpretation were addressed using the tabular module, which created a table linking each verse number to its interpretation. The last module was based on building an ontology to retrieve the answer. Heidaria *et al.* (2014) implemented this system using the GATE framework.

However, this system had many flaws. It faced issues with Persian preprocessing tools, such as tokenisers. Additionally, the coverage of the ontology was limited. Moreover, the Persian NER tool had a weak performance. The system also suffered from limitations in the number of corpus questions for some question types. Finally, this system only answered Persian questions about the Quran.

Sherkat & Farhoodi (2014) suggested improving the Quranjooy system by using a hybrid approach to identify the type of question through classification based on machine learning and rule-based methods. The proposed method's precision was approximately 56%.

Sheker *et al.* (2016) developed two Fatwa QA systems: an ontology-based system and a synonym-set system. Muslims typically ask Islamic scholars or sheikhs for their opinions on questions regarding various daily issues faced by the questioner; the answer is known as a fatwa . The two knowledge resources created for these systems are a corpus and an ontology. First, a corpus of 1,094 QA pairs in the area of prayer-related fatwas was constructed, sourced from Ibn Baz's book (Ibn Baz *et al.*, 2003). Second, an ontology was built based on this corpus, consisting of 13 main classes, 134 sub-classes, 19 relations and 511 annotations. The statistical information from the corpus defined the main and sub-concepts in the ontology. Moreover, experts in Arabic had determined the relationships between these concepts.

The architecture of an ontology-based system and a synonym-set system involves three stages: question pre- processing, question analysis and question expansion. Encoding and normalisation operations are implemented to pre-process the question. The goal of the question analysis phase is to return corpus questions that are similar to the user's question. First, the similarity measures using the cosine and Jaccard methods produced two lists of candidate questions. Afterward, the candidate questions were ranked based on their frequency among the lists. Finally, the most relevant question was retrieved from the dataset, and its answer was displayed to the user. The architecture of the two systems was the same, with a difference only in the question expansion phase. The question expansion reformulated the query by adding a synonym. The first system used synonyms found in a new ontology, which was a specific-domain ontology, while the second system used WordNet synonyms, which was an open-domain ontology. Thirty questions, as proposed by experts, were used to evaluate the systems. The performance of the ontology-based system was better than that of the synonym-set system. The first system achieved 91% on the F-measure, 90% recall, and 92% precision, while the second obtained 65%, 59%, and 72% on the same metrics, respectively.

However, the ontology had limited concepts, and Arabic WordNet had weaknesses that affected performance. Additionally, the system could only answer a limited number of questions about prayer. The answer was a fatwa based solely on the words of scholars without evidence from verses or hadiths.

Hamoud & Atwell (2016a,b, 2017) proposed a system to answer Arabic and English Quran questions using a closed-domain knowledge base. This system consisted of two components: a question–answer pairs corpus and a QA model. The 1,500 question–answer pairs were collected from different resources. This knowledge base included various types of questions, such as factual, hypothetical, how-to, definitions and both long and short questions. To enhance data quality, the researcher cleaned the data. Next, they represented the corpus in a comma-separated value (CSV) format. After that, the system's performance was improved by applying data redundancy techniques, such as paraphrasing questions in different ways and contexts. Finally, they loaded the latest version of the data into the Nooj and Weka tools for analysis. The model of this system was based on accepting questions from users through an interface. Next, it analysed a question by implementing pre-processing operations. It then matched the preprocessed question with dataset questions using a text-based approach. After that, it scored and ranked the matched questions to select the best match using a similarity measure. Lastly, it presented the answer to the users. They used the Python Natural Language Toolkit (NLTK) to implement this module. They evaluated the system using 63 English and 71 Arabic questions gathered from Muslim students at a university. The Arabic version of this system achieved 79% precision and 76% recall, while the English version obtained 75% precision and 73% recall.

However, as mentioned previously, this system used only the text-based approach when retrieving answers, which has many inherent limitations. It answered only Quran-related questions. The performance of this type of system is affected by two factors: the quality and quantity of the data in the corpus. Unfortunately, there were many flaws in the corpus used. The first limitation was that a high percentage of questions in the corpus were unreliable because they had not been checked and examined by a religious scholar, especially the data extracted from previous research. In addition, the references for each question and answer were not provided, and many answers lacked supporting evidence from the Quran.

Adany *et al.* (2017) recommended a QA system with two stages for the Quran domain. The first stage involved constructing a corpus, while the second stage focused on designing two prototypes for QA systems based on the corpus. The scope of the corpus included the Al-Baqarah and Al-Fatiha chapters. The researchers used two methods to create the corpus: gathering question–answer pairs from Islamic websites and generating questions manually from the Quran verses. This corpus consisted of 263 Quran questions, with one question potentially having multiple verses as answers. It was divided into two categories: the first was used as a gold standard evaluation dataset, and the second was used as the knowledge base in the study.

The use of the first prototype involved the following steps:

(a) Accept the user question and preprocess it by applying a tokeniser and removing diacritics and stop words.

(b) Match the user question with the question–answer corpus.

(c) If the question exists in the corpus, then:
    The answer will be retrieved and displayed to the user.
    Exit.

(d) If the question does not exist in the corpus, then:
    The search process will be applied to find the answer from the Quran text.

    i. If the answer is located in the Quran text, then:
        The answer will be retrieved and displayed to the user.
        The question and answer will be added to the corpus.
        Exit.

    ii. If the answer is not found in the Quran text:
        The question will be added to a specialised file in the corpus.
        The specialised file will be checked, and the question will be answered by administrators and scholars.
        Exit.

The prototype was implemented using Java Standard Edition version 8. The developer also suggested another prototype based on the corpus as a knowledge base, but it was not implemented. The use of the second prototype involved the following steps:

- The model accepts the question from the user.
- If the question is in the question–answer corpus, the model will return the answer and display it.
- If the question is not in the question–answer corpus, the system will execute the following series of steps to determine the answer:
  - The question will be pre-processed using a tokeniser, classifier and remover.
  - The question will be expanded using a thesaurus, dictionary and corpus.
  - The new question will be displayed to the user for appropriate wording.
  - The search process with the new question will be conducted using the question–answer corpus or the Internet.
  - The answer will be extracted and displayed to the user.
  - If the user approves the answer, both the question and the answer will be added to the corpus.

Nevertheless, this system suffers from many shortcomings. Its coverage is limited to answering 263 questions from the Surah Al-Fatihah and Surah Al-Baqarah. The answer provided is the entire Quran verse, rather than a specific answer or explanation. Additionally, this system fails to answer questions that are similar in meaning to the corpus questions but use different terms. A religious expert did not comprehensively verify the answers provided by this system.

Despite the importance of having a system to accurately answer all kinds of religious questions, the amount of research in this field is very limited. The majority of existing studies focus on answering Quranic questions, with only one addressing fatwa questions. Additionally, no study has answered questions related to the field of Hadith.

2. **Knowledge Base System using Deep Learning Models**

   Typically, understanding the meaning of a question and providing an answer are done using one of the previously mentioned methods, such as ontology and WordNet. The problem with these knowledge sources is that they are expensive to create and are often deficient because they do not cover all concepts in the field. Recently, DL modules have emerged as alternatives. DL models aim to reach a level of understanding comparable to that of humans. These models can understand and answer various questions (Altammami *et al.*, 2021).

   The most recent research in NLP tasks tends to use the corpus to train PLMs for building QA models because they have proven effective (Malhas & Elsayed, 2022). Additionally, most research has recently shifted towards using LLMs (Katz *et al.*, 2023). I will discuss this topic in detail in Chapter 5, Chapter 6, Chapter 7 and Chapter 8.

## 3.5 Conclusion

This chapter provides an in-depth and sequential overview of the various developments and methods used over the years in NLP to build a QA system based on the Holy Quran and Hadith. The religious question-answering system can be divided into retrieval-based systems and knowledge-based systems.

Retrieval-based systems extract answers from texts, relying entirely on NLP tools and techniques during the question analysis phase. However, these tools still face challenges and have many limitations, making it difficult for such systems to answer all types of questions.

Since the question and answer may contain different words with the same meaning, some systems of this type utilize external sources, such as ontologies, to retrieve all answers related to a concept in the ontology that corresponds to the word in the question. However, a major drawback of these systems is their high cost, as they require the development of an ontology for each domain.

Additionally, all existing ontologies are incomplete and do not encompass all concepts in the field.

The knowledge-based systems focuses on building a dataset containing a large number of questions and answers. Systems of this type are divided into two categories.

The first category utilizes NLP tools to extract a question similar to the user's query from the dataset and then displays it along with its answer. However, this approach is limited by the performance of NLP tools, which have several shortcomings

The second category involves using the dataset for training DL models, which is the current trend, or as a knowledge base for LLMs. This was studied in the following chapters.

# Chapter 4

# Developing the Quran and Hadith Question–Answer Corpus

## 4.1 Introduction

This chapter surveys and reviews datasets of Islamic religious question–answer pairs. The findings indicate that there are many defects in the collections related to the Quran, while no available collection was found related to the Hadith. Therefore, this chapter presents QUQA, the most extensive reusable Quran question–answer collection, by integrating existing datasets and enlarging them using different resources and challenging questions. This dataset covers many questions and more verses, with the questions in MSA and the answers from Quran verses in CA. This chapter also introduces HAQA, the first reusable Arabic hadith question–answer corpus, by collecting data from various expert sources. These two datasets are available to the research community, which will positively impact research on Islamic QA. Furthermore, these datasets enrich the resources for the Arabic language, which suffers from a shortage of AI datasets and challenges related to studying the nature and understanding of CA texts. Finally, all available Automatically Generating Questions (AGQ) tools were surveyed and evaluated to determine their potential to expand the Quran dataset. This chapter is based on my publications (Alnefaie *et al.*, 2022a, 2023c,d).

## 4.2 Religion Corpora Related Work

The Internet contains vast amounts of data written using various software programs using assorted templates and diverse formats. Gathering this data into a single corpus was essential because it forms the foundational block in many areas of data mining, such as classification and QA. The characteristics of a corpus, such as its size, affect system performance. As the amount of data increases, the performance of a system using it generally improves. For example, the performance of object classification algorithms can be enhanced more by increasing the data size than by improving the learning algorithms (Sapp *et al.*, 2008).

Several religious corpora have been built. For example, Mohammed *et al.* (2022) created an English Islamic Articles Dataset (EIAD), which contained ten thousand articles collected from three websites: IslamReligion (1,550 articles) ,IslamQA (8,292 articles), and New Muslims (275 articles). The target users of this dataset were new Muslims or individuals who wanted to convert to Islam. It is organised into 15 categories, each containing related folders of topics, with each topic encompassing relevant articles. The metadata for the corpus included information about the articles, such as publishing date, author, rating and description. Nevertheless, it was a collection of articles from a limited number of sources, and its topics were relevant only to new Muslims or those willing to convert to Islam; in addition, it is currently unavailable.

Several question–answer corpora have been developed for religious QA systems and can be classified into two categories: question–answer pairs datasets and question-passage-answer triplets datasets. Both are described below.

### 4.2.1 Question–Answer Pair Datasets

Saeedi *et al.* (2014) developed the Quranjooy corpus, a collection of Quran question–answer pairs in Persian. The question–answer dataset was created by crawling many credible websites. First, a list of found sites was created. Quranic experts then selected authorised sources. The total number of collected question–answer pairs was approximately 115,000, formatted in XML. After that, a data cleaning process was applied, and questions with grammatical or editing issues were excluded. The final corpus contained 6,000 questions covering various topics, such as characteristics, places, personalities, numbers, creatures, behaviours , events and dates. Nevertheless, it did not cover all areas of the Quran and was written in Persian.

Sheker *et al.* (2016) proposed a corpus of 1,094 question–answer pairs for the prayer fatwas domain, which was gathered from Ibn Othaimeen's fatawa book (Ibn Baz *et al.*, 2003). The corpus specialised in questions solely related to the subject of prayer and was characterised by its small size. Moreover, the answers reflected scholarly opinions and beliefs without any supporting evidence from the Quran or Hadith.

Hamdelsayed & Atwell (2016) and Adany *et al.* (2017) developed a question–answer pairs dataset regarding Surah Al-Fatiha and Surah Al-Baqarah in the Holy Quran. This corpus was used for two purposes in the research's QA prototypes: as a knowledge base for the system and as a gold standard evaluation dataset. Building this corpus involved several steps: the researchers collected the question–answer pairs from websites, extracted the questions by reading the Quran text and integrated the two sets into a single text file, an Access file, and a Microsoft Excel file.

They elicited 47 questions from Islamic websites, but the gathered text suffered from many problems, such as diacritics and English characters. Therefore, the text required some cleaning. Moreover, the researchers extracted the appropriate questions from the verses. One question could have multiple answers. The total number of questions generated using this approach was 215. These questions were reviewed and validated by Islamic scholars from Gabrah College. The combined dataset contained 263 questions. The data in the file were arranged according to the verse number to detect duplicate questions. The final file contained five columns: the question, answer, verse number, surah name and a column for Abdullah Yusuf Ali's English translation of the verse.

However, the corpus covered only Surah Al-Fatihah and Surah Al-Baqarah, and the number of questions was very limited. It is currently unavailable.

Hamoud & Atwell (2016a) developed a dataset of question–answer pairs on the Quran by following these steps: the researchers manually gathered 1,500 question–answer pairs from Islamic websites, previous research and experts at the Mecca Holy Mosque and then merged them into one knowledge base. Next, the corpus was cleaned and formatted appropriately. They then analysed the dataset using the WEKA and Nooj tools. Finally, it was used as knowledge for a QA search tool.

The eight websites used in this knowledge base were turntoislam [1], Islamic Knowledge/Come towards Islam [2], All-Quran [3], The Siasat [4], SULTAN ISLAMIC LINKS [5], Islamic question and answer [6], Sheikh Dr. Mohammad Al Arifi official forum [7], and the Sheikh Hussaballa forum [8]. They used examples mentioned in the studies by Gusmita *et al.* (2014), Abdelnasser *et al.* (2014), Shmeisani *et al.* (2014) and Hamdelsayed & Atwell (2016). One of the services available in the Holy Mosque is the ability to ask Islamic experts questions. The researchers communicated with Muslims who returned from Mecca to collect the questions posed to the experts and the answers they received.

However, this corpus contained only 1,500 question–answer pairs. There was no evidence from the Quran in many of the answers, and the question–answer pairs in this corpus lacked references.

Neamah & Saad (2017) built a corpus of 12 questions to evaluate their Hadith QA system. They asked 15 students from Universiti Kebangsaan Malaysia (UKM) universities to create a list of questions.

Abdi *et al.* (2020) constructed an Arabic corpus of 3,825 question–answer pairs about Hadith from the Sahih al-Bukhari collection. Two human experts built the corpus manually by (1) reading all the hadiths in order, (2) generating questions for each hadith, (3) removing duplicate questions, and (4) linking each question with the corresponding hadith. There were different types of questions, but the most significant percentage was for "WH" questions.

Maraoui *et al.* (2021) collected a dataset of 100 question–answer pairs from online forums and native Arabic speakers. The distribution of this dataset was

[1] http://turntoislam.com/community/threads/100-questions-on-quran.10052

[2] https://islamicknowledge2all.wordpress.com/2011/10/30/question-and-answers-about-quran-3/

[3] http://www.allquran.com/islamic_material/frequently_asked_questions.html.

[4] http://www.siasat.com/english/news/questionsanswers-about-holy-quran?page=0%2C0

[5] http://www.sultan.org/.

[6] http://islamqa.info/en/

[7] http://www.3refe.com/vb/

[8] http://hassabala.yoo7.com/t714-topic?highlight=500+%D3%C4%C7%E1.

as follows: 13 questions about the Tafsir topic, 33 questions about the Hadith topic and 54 questions about the Hadith narrator profile. Nevertheless, it was an unauthenticated and small corpus.

Munshi *et al.* (2022) constructed an Arabic fatwa dataset with 850,000 records. Each record contains a question, answer, Fatwa topic and publication date. The dataset focused on questions and answers from social media channels, unlike existing datasets that focused on Quran and Hadith texts. Typically, users posted their questions, and highly qualified experts provided responses. They gathered the questions from various countries, backgrounds and dialects of Arabic . The resources could be classified as authentic-government sources, such as Al-ifta-SA [1] and Dar-al-ifta-EG [2] , as well as untrusted websites like fatawapedia [3], Islamweb [4], Islamway [5], binothaimeen [6], AlFawzan [7], Islamqa [8], and binbaz [9]. Some websites contained only articles, treating the title as a question and the article as an answer. A sample is shown in Figure 4.1. Nevertheless, these fatwas lacked evidence from the Quran or Hadith.

| | | |
|---|---|---|
| **Question** | Arabic Egyptian accent | س: بسبب العلاج ومواعيد مش هفطر فى رمضان غير أقل من 3 ساعات وباقى اليوم صايمه هو انا كدة ينفع مصومش؟! |
| **Question** | English | Because of the treatment, I will break my fast in Ramadan for less than three hours a day and fast for the rest of the day. Is that permissible? |
| **Answer** | Arabic Egyptian accent | ج: إذا كان يتعبك الصوم فلك رخصة في الفطر في هذه الحالة |
| **Answer** | English | If you are tired because of fasting, then, in this case, you have permission to break the fast |

Figure 4.1: Part of the fatwa dataset record.

[1] https://www.alifta.gov.sa
[2] https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&1&Home=1
[3] http://fatawapedia.com/
[4] https://www.islamweb.net/ar/
[5] https://ar.islamway.net/fatawa/source/
[6] https://binothaimeen.net/site
[7] https://www.alfawzan.af.org.sa
[8] https://islamqa.info/
[9] https://binbaz.org.sa/fatwas/kind/1

All previous datasets are unavailable for evaluating QA systems in the religious domain. The current research uses a different dataset created by the authors. As a result, it is difficult to compare performance between systems. To the best of our knowledge, only two datasets have recently been developed for the Arabic Quran and are available to the public. They are described below.

Alqahtani (2019) built a corpus of 2,224 question–answer pairs related to the Holy Quran to test and evaluate his QA system, but the available version only contained 1,224 question–answer pairs due to copyright issues. This collection was called the Annotated Corpus of Arabic Al-Quran Question and Answer (AQQAC). The construction of this collection went through four stages. First, the data were gathered from the book 1,000 Questions and Answers on the Quran, written by Islamic scholar Ashour (2002), and from a website named 'Islam – the Quran and Tafseer' [1]. Second, the data were preprocessed and cleaned using regular expression patterns to remove unwanted elements, such as non-letter characters and page numbers. Third, the question–answer pairs were entered into an Excel spreadsheet. Finally, the question–answer pairs were annotated with descriptive information. This information included details related to the questions, such as the ID, type, topic and ontology concepts, as well as the answer's location in the Quran (chapter and verse numbers) and the reference sources for these questions and answers.

However, there were deficiencies in this dataset that affected its use for evaluating QA systems. Many of the answers were merely interpretations of the verses and did not contain any Quranic verses. Other answers contained verses mixed with interpretations, and the last type of answer was written in MSA and was supported by evidence from the Quran, in which the evidence was an entire verse. The problem was that these answers had the same meanings as the Quranic words but used different terms. I had to extract the answer span from the full verse of the Quran. After filtering out cases with no Quranic verses, the dataset contained only 473 out of the original 1,224 question–answer pairs (Aftab & Malik, 2022). Additionally, the dataset lacked diversity because the available version was built from only one source and a small corpus.

[1] http://islamqt.com/

Malhas & Elsayed (2020) constructed a gold standard Arabic corpus of 207 questions about the Quran, called AyaTEC, to unify the QA systems' assessment process. There were 1,762 verses in the answers, and each answer to the question included one or more verses. This corpus covered 11 topics: provisions of Islam, stories of prophets, former nations, the unseen, the universe and God's creations, worship, jihad, battles and wars, faith in God and believers, Prophet Muhammad, humanity, linguistics of the Quran and others. The researchers collected the questions from users directly and from multiple other sources, including books, YouTube videos and previous studies, such as Hakkoum & Raghay (2016) and Abdelnasser *et al.* (2014). Two freelancers from UpWork[1] answered the questions. Subsequently, three religious scholars reviewed the answers to verify their validity. Nevertheless, the dataset size was too small, with only 207 questions, and there was no variety in the questions, as half were on only two topics. Additionally, this dataset covered only 1,762 verses, while the total number of verses in the Quran is 6,236.

### 4.2.2 Question–Passage–Answer Triplet Datasets

A question–passage–answer triplets dataset is an MRC dataset, which is typically used to help a person better understand a text by asking questions about it and measuring their comprehension. Each record contained a question, a passage from the Quran and an answer extracted from that passage.

Four studies have developed this type of dataset for the Arabic Quran and three approaches have been used. The descriptions appear below.

1. Some researchers only changed the structure of an available dataset from pairs (question and answer) to triplets (question, passage and answer).

2. Some researchers modified the structure of available question–answer pair datasets to fit the triplet structure (question, passage and answer) and added new questions similar to the existing ones.

3. Some researchers expanded a dataset by formulating questions in several ways.

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

| | % | # of the Pairs | # of the Triplets | Passages | Questions |
|---|---|---|---|---|---|
| **Training** | 65% | 710 | 861 | 468 | 118 |
| **Development** | 10% | 109 | 128 | 101 | 17 |
| **Test** | 25% | 274 | 348 | 256 | 34 |
| **All** | 100% | 1,093 | 1,337 | 825 | 169 |

Table 4.1: The statistics for the number of records in the QRCD.

1. **Restructuring the Existing Datasets**

   The Quranic Reading Comprehension Dataset (QRCD), a publicly accessible dataset[1] proposed by Malhas *et al.* (2022), is an MRC version of AyaTEC. Tanzil's project[2] was the source of the Quranic text. It contains 1,093 question-passage pairs written in the JSON file format, but because the questions could have more than one answer, the triplets become 1,337 question-passage-answer sets, as shown in Table 4.1 and Figure 4.2. A passage may appear more than once but with different questions; similarly, a question may appear more than once but with a different passage. The triplets were classified into training, development and test sets.

   However, the number of non-repetitive questions was very small, at only 169. Moreover, not all correct answers were added to the dataset. For example, the gold answer to the question متى يحل الأسلام دم الشخص, which translates to 'When does Islam allow the blood of a person?' is قاتلو في سبيل الله الذين يقاتلونكم, meaning 'Fight in the way of Allah those who fight you'. In addition, there was another correct answer in the passage فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم, meaning 'So whoever has assaulted you, then assault him in the same way that he has assaulted you'. The meaning of this other correct answer is the same as the gold answer, according to bin Jarir al Tabari (2001), as shown in Figure 4.3 (Alsaleh *et al.*, 2022).

[1] https://gitlab.com/bigirqu/quranqa
[2] https://tanzil.net/docs/tanzil_project

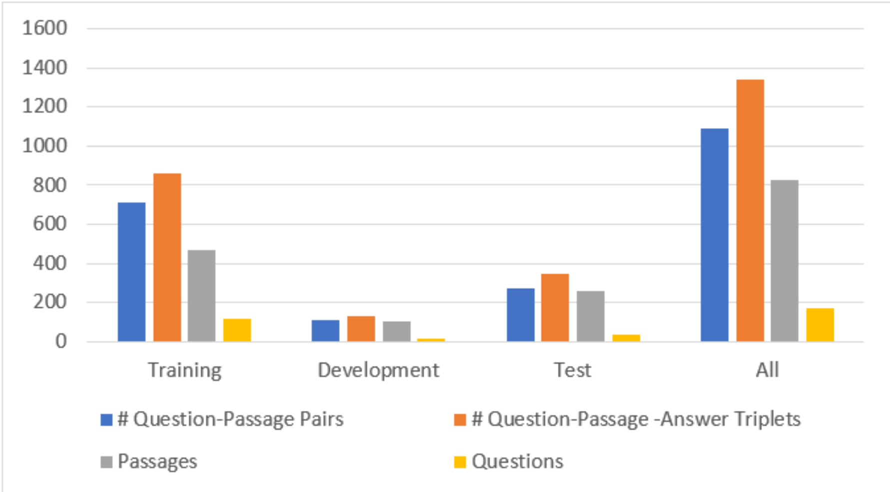
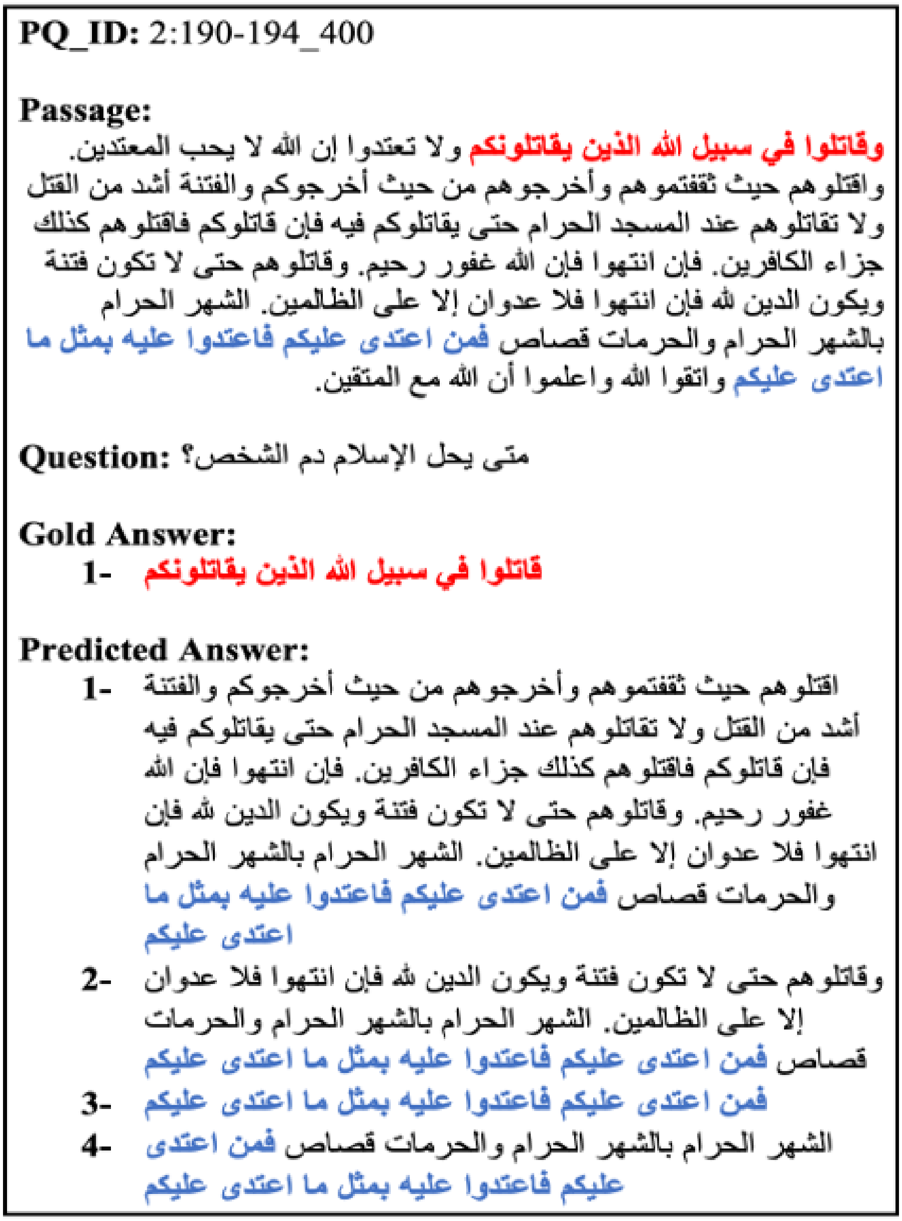



Figure 4.2: The records distribution of the QRCD.

Aftab & Malik (2022) extracted 473 question–answer pairs from the AQQAC to augment the QRCD dataset and enhance their QA system. The structure of AQQAC was different from QRCD, as the AQQAC records contained the question and the answer, which was a verse or sequential verses from the Quran. In contrast, QRCD records contained the question, a passage (a verse or sequential verses) and an answer that was a partial sentence or word from the passage. Therefore, they restructured AQQAC to match the QRCD structure by adding a context column. For each question, they filled this column through the following steps: locating the verse in the answer within the Quran using the Tanzil text; extracting the verse preceding the answer's verse, the answer's verse and the verse after it; and finally, adding these verses to the context column. Nevertheless, this dataset is currently unavailable.

2. **Restructure the Existing Datasets and Added New Questions**

   Wasfey *et al.* (2022) constructed the AQAQ corpus based on the AQQAC dataset, which consisted of 732 question-passage-answer triplets. The construction process involved several stages. First, they filtered 500 questions, keeping only the verses as answers and deleting any other answers. Next, a Quranic scholar answered all questions that initially lacked answers and



**PQ_ID:** 2:190-194_400

**Passage:**
وقاتلوا في سبيل الله الذين يقاتلونكم ولا تعتدوا إن الله لا يحب المعتدين.
واقتلوهم حيث ثقفتموهم وأخرجوهم من حيث أخرجوكم والفتنة أشد من القتل
ولا تقاتلوهم عند المسجد الحرام حتى يقاتلوكم فيه فإن قاتلوكم فاقتلوهم كذلك
جزاء الكافرين. فإن انتهوا فإن الله غفور رحيم. وقاتلوهم حتى لا تكون فتنة
ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام
بالشهر الحرام والحرمات قصاص فمن اعتدى عليكم فاعتدوا عليه بمثل ما
اعتدى عليكم واتقوا الله واعلموا أن الله مع المتقين.

**Question:** متى يحل الإسلام دم الشخص؟

**Gold Answer:**
1- قاتلوا في سبيل الله الذين يقاتلونكم

**Predicted Answer:**
1- اقتلوهم حيث ثقفتموهم وأخرجوهم من حيث أخرجوكم والفتنة أشد من القتل ولا تقاتلوهم عند المسجد الحرام حتى يقاتلوكم فيه فإن قاتلوكم فاقتلوهم كذلك جزاء الكافرين. فإن انتهوا فإن الله غفور رحيم. وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام بالشهر الحرام والحرمات قصاص فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم
2- وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام بالشهر الحرام والحرمات قصاص فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم
3- فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم
4- الشهر الحرام بالشهر الحرام والحرمات قصاص فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم

Figure 4.3: An example of another correct answer can be added to the dataset.

added the appropriate passages. Finally, similar questions were included to increase the size of the dataset to 732 question–passage–answer triplets. Although it is shown as being available [1], it could not be opened at the time of this writing.

3. **Reformulating the Existing Questions**

   Ahmed *et al.* (2022) paraphrased the questions to augment the training and development sets of the QRCD by reordering the words and replacing them with synonyms. This available augmentation dataset[2] was used to fine-tune the model so it could find the correct answers to different question formulations in the testing phase. Its size is approximately 657 question-passage-answer triplets.

### 4.2.3 Religious Question–Answer Corpora Evaluation Criteria

This study aimed to survey and evaluate the existing corpus of question–answer pairs about religious texts, which included more than 200 questions. This effort sought to identify flaws in current knowledge bases to develop a larger, improved collection by reusing the available datasets. Several criteria were used to evaluate the corpora , such as the scope, size and purpose of their construction. Thirteen criteria were employed in this study to assess the existing question–answer pair datasets. Some of these measures were adapted from Alrehaili & Atwell (2014) to evaluate Quran ontologies. The details of these criteria are listed below.

1. Resource of the Question:

   A. Books
   B. Websites
   C. Religion experts
   D. Previous studies
   E. Users directly
   F. The author who extracted the questions from the text

[1]https://github.com/EmanElrefai/Quran_QA/tree/main/datasets
[2]https://github.com/motazsaad/Quran-QA

2. Resource of the Answer:

   A. Books B. Websites

   C. Religion experts D. Previous studies

   E. The author who extracted the answers from the text

3. Purpose of Creating this Corpus: The authors created this dataset to use as:

   A. Gold dataset to evaluate the question–answer system

   B. A knowledge base for the question–answer system

4. Questions Type: The types of questions in this corpus

   A. Factoid B. Definition

   C. List

   D. Other: Yes/No, Facts, Arguments, relation, etc.

5. Answer Type: The nature of answers in this corpus includes

   A. Plain Texts B. Verse

   C. Verse number and Surah name D. Hadith

   E. Explanation from Tafsir books F. Part of the verse

6. The Corpus Language:

   A. Arabic B. English

   C. Persian

7. Corpus Size: The number of question–answer pairs in this corpus.

8. Scope of the Corpus:

   A. Quran B. Hadith

   C. Tafsirs: Descriptions of the Quran D. Fatwas

9. The Data Coverage Area:

   A. The whole book B. Parts of the book

10. Topics Coverage:
11. Corpus Availability: Is this corpus available to the users?

    A. Available B. Not available

12. Corpus Formats:

    A. Text B. CSV

    C. XML D. Access Database

    E. JSON Lines file

13. Validation Approaches: used to validate the corpus:

    A. Reviewed by an Islamic scholar

    B. An Islamic scholar answered the questions in a book or website

    C. None

The previous thirteen evaluation criteria are used to compare the religious question–answer pairs dataset. Tables 4.2 and 4.3 show the comparison results. Due to space constraints, datasets are named either as assigned by the authors or, if unnamed, I assign them using the first author's name.

| Criteria | Saeedi | Sheker | Hamdelsayed | Hamoud | AQQAC | AyaTEC |
|---|---|---|---|---|---|---|
| **1.Resource of the Question:** | B | A | F and B | B, C, and D | A and B | A, B, C, and D |
| **2.Resource of the Answer:** | B | A | E and B | B, C, and D | A and B | C |
| **3.Purpose of Creating this Corpus** | B | B | A and B | B | A | A |
| **4.Questions Type** | - | - | - | - | A, B, C, and D | A, B, C, and D |
| **5.Answer Type** | A | A | B and C | A | A, B, C and E | B, C, and F |
| **6.The Corpus Language** | C | A | A | A and B | A | A |
| **7.Corpus Size** | 6000 Q&A pairs | 1094 Q&A pairs | 263 Q | 1500 Q&A pairs | 2224 Q&A pairs | 207 questions |
| **8.Scope of the Corpus** | A | D | A | A | A and C | A |
| **9.The Data Domain Coverage** | A | A | B | A | A | A |
| **10.Topics Coverage** | 30 topics | Prayer | - | - | - | 11 topics |
| **11.Corpus Availability** | B | B | B | B | Partly A | A |
| **12.Corpus Formats** | C | - | A, B, and D | B | B | A and C |
| **13.Validation Approaches** | A | B | Part of it uses A | Part of it uses A | Part of it use B | A |

Table 4.2: Comparing the religion question–answer corpus part 1.

| Criteria | Abdi | QRCD | Aftab | Wasfey | Munshi |
|---|---|---|---|---|---|
| **1.Resource of the Question:** | C | A, B, C, and D | D | C and D | B |
| **2.Resource of the Answer:** | C | C | D | C and D | B |
| **3.Purpose of Creating this Corpus** | A | A and B | B | B | B |
| **4.Questions Type** | - | A | - | - | - |
| **5.Answer Type** | D | B, C, and F | B, C, and F | B, C, and F | A |
| **6.The Corpus Language** | A | A | A | A | A |
| **7.Corpus Size** | 3825 Q&A pairs | 1,337 triplets | 473 triplets | 732 triplets | 850K Q&A pairs |
| **8.Scope of the Corpus** | B | A | A | A | D |
| **9.The Data Domain Coverage** | A | A | A | A | A |
| **10.Topics Coverage** | - | - | - | - | - |
| **11.Corpus Availability** | B | A | B | A | B |
| **12.Corpus Formats** | - | E | E | E | - |
| **13.Validation Approaches** | B | A | B | A and B | Part of it use B |

Table 4.3: Comparing the religion question–answer corpus part 2.

The preceding discussion can be summarised as follows:

- The creation process of question–passage–answer datasets for the Quran is in its early stages and requires significant work, as the available dataset size has been very small and the number of questions does not exceed hundreds.
- There is an absence of a gold standard dataset used to measure the performance of QA systems in the Hadith area. Therefore, researchers have had to create their own datasets to evaluate the performance of their systems by collecting question–answer pairs from various websites or books or creating them. As a result, each system used different sizes and types of question–answer pairs, making it difficult to compare the performance of these systems fairly.

## 4.3 Questions Generator Web-Based Tools Related Work

The process of automatically generating questions (AGQ) is a task in NLP. This process aims to generate question–answer pairs from various inputs, such as knowledge bases, text, and images, as shown in Figure 4.4. This thesis focuses on generating questions from text, which has become very popular recently due to its benefits in many wide-ranging applications, such as intelligent tutoring systems, QA, MRC, and customer service chatbots.

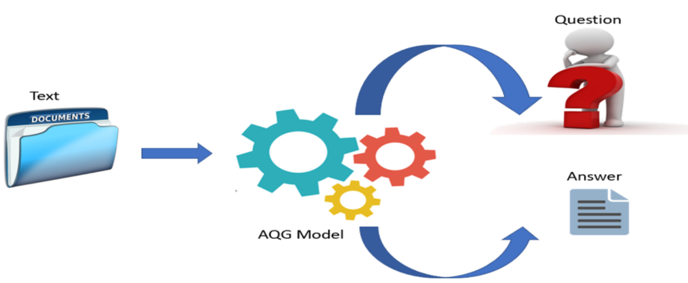



Figure 4.4: The structure of the automatic question generation tool.

There are many tools available on the web that can be used to generate questions. Question generation web-based tools can produce different question types for a paragraph, such as true/false, multiple-choice, and short-answer questions. These tools can be classified into short-answer question generator tools and other question type generator tools. A comparison of these web service tools is shown in Tables 4.4, 4.5, and 4.6. The criteria used for the comparison are: (1) the size of the text allowed to be entered, (2) the question types that can be generated, (3) the number of questions that can be generated using the tool, (4) whether the tool is free or requires a subscription, (5) the language in which the text can be entered, (6) whether the tool is automatic or requires user intervention, (7) the time the tool requires to complete its task, and (8) the algorithm used by this tool to generate questions. These tools are discussed below.

| **Tools** | **Input Size** | **Types of Questions** | **Number of Questions** | **Free** | **Language** | **Automatic Tool** | **Duration** | **Algorithms** |
|---|---|---|---|---|---|---|---|---|
| ExploreAI Question Generation demo | 1000 character | Short Answer | 1 to 4 | yes | English | yes | 75 Second | Text-To-Text Transfer Transformer (T5) |
| Cathoven QG | 500 words | Short Answer | 1 to 20 | yes | English | yes | 60 Second | Text-To-Text generation model |
| Questgen QG | 50 - 500 words | MCQ Yes/No Short answer | 0 to 18 | application and API: no,open source NLP library: yes | English | yes | 7 Second | GPT-3, GPT-2, T5, and BERT |
| Lumos learning QG | at least 2000 characters | Short Answer | 0 - 32 | yes | English | yes | 40 Second | - |

Table 4.4: Comparison of web services question generation tools part 1.

| **Tools** | **Input Size** | **Types of Questions** | **Number of Questions** | **Free** | **Language** | **Automatic Tool** | **Duration** | **Algorithms** |
|---|---|---|---|---|---|---|---|---|
| ParaQG (Kumar *et al.*, 2019) | - | Short Answer | - | yes | English | no-need user intervention to determine possible answers | - | a sequence-to-sequence model, global sparse-max attention, copy mechanism, and dynamic dictionary |
| QuestionAid QG | 1500 characters | Short Answer | 5 | no | 24 languges | yes | 38 Second | - |
| PrepAI QG | No limit | MCQ ,Fill the blank,Short Answer ,True / False | Easy MCQ 36, Medium MCQ 51, Hard MCQ 2, Fill the blank 18, Short Answer 0, True / False 36 | no | English | yes | 1 minute and 29 second | - |

Table 4.5: Comparison of web services question generation tools part 2.

| Tools | Input Size | Types of Questions | Number of Questions | Free | Language | Automatic Tool | Duration | Algorithms |
|---|---|---|---|---|---|---|---|---|
| Quillionz QG | 300-3000 words | MCQ, Fill the blank ,Short Answer | MCQ 26, Fill the blank 18, Short Answer 10 | no | English | yes | 1 minute and 54 second | - |
| Automatic QG | - | MCQ, Descriptive, True / False | MCQ 31, Descriptive 2, True / False 2 | no | English | yes | - | - |
| QG API | No limit | Fill the blank | 10 | yes | English | yes | 7 Second | - |
| pedagogic tool (Vu & Blake, 2021) | - | Closed-ended question Tag question Open-ended question | - | yes | English | yes | - | The rules-based Approach for the syntactic |

Table 4.6: Comparison of web services question generation tools part 3.

### 4.3.1 Short Answer Questions Generator Tools

The Joint Information Systems Committee (JISC) developed a question generator tool called the ExploreAI Question Generation (QG) demo [1]. This demo allows users to enter a paragraph limited to 1,000 characters and uses the Text-to-Text Transfer Transformer (T5) to automatically determine answers from the text and generate the appropriate questions for these answers. The generated questions are then displayed. Additionally, it allows the user to answer questions and then either display the correct answers or show the correct answers directly. Between one and four questions are generated for each paragraph, and the question generation process takes 75 seconds. One disadvantage of this system is that any grammatical or factual errors in the text will appear in both the answers and the questions. This code is available [2].

Cathoven's team built a question creator tool called the Cathoven QG [3]. Users must enter the text and specify the number of questions they want to generate. This tool can accommodate a text passage of up to 500 words as input. There are three options available regarding the number of questions to be generated: 'few,' 'many,' and 'tonnes.' Then, using the text-to-text generation NLP model, the encoder-decoder structure is applied to analyze the details of the text and quickly generate the questions. It typically takes about a minute to produce questions, which are then displayed. When the 'many' option is chosen, between one and 20 questions will be generated. The user can answer the questions, and the model will indicate whether the answers are correct or incorrect. This model is smart enough to consider the user's answer valid if it has the same meaning as the correct answer. The user can also view the correct answers. The questions and answers can be exported as TXT files. A sample of the generated questions using Cathoven QG is shown in Figure 4.5.

The Questgen QG tool [4] is built on a combination of computational linguistics and advanced AI algorithms. Many transformer models are used, including GPT-3, GPT-2, T5, and BERT. It is available as an application, an API, and through

[1] https://exploreai.jisc.ac.uk/tool/question-generation
[2] https://github.com/patil-suraj/question_generation
[3] https://hub.cathoven.com/?scene=generator
[4] https://questgen.ai/

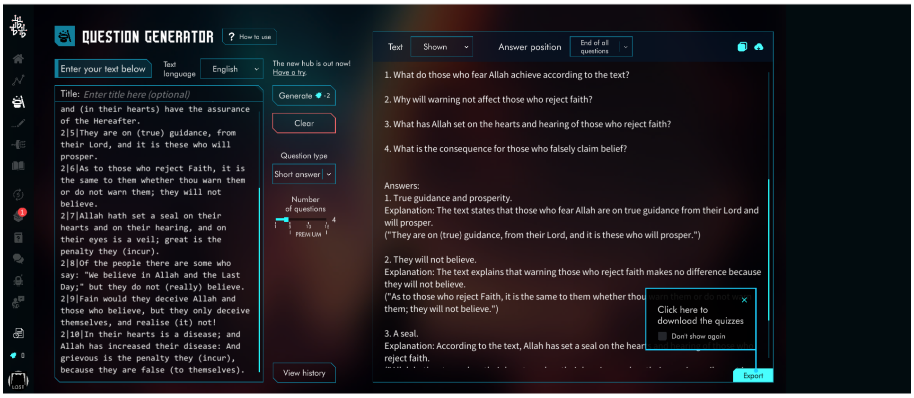



Figure 4.5: The sample of the generated questions using Cathoven QG.

an open-source NLP library. There is a free trial application that allows only 15 free attempts. To gain access to unlimited tries and additional features, a subscription to the application and the API is required. The Questgen QG open-source library is less accurate than the application and the API. Three types of questions can be generated: multiple-choice questions (MCQs), Boolean questions (yes/no), and short-answer questions. Users need to input text ranging from 50 to 500 words and select the type of questions, after which the questions and answers will appear within seven seconds.

Lumos Learning is an EdTech platform that aims to improve the learning process using digital solutions [1]. One of its most important features is the ability to generate assessments for reading comprehension, mathematics, and language and grammar using the Lumos Learning QG Tool. After a passage of at least 2,000 characters has been pasted into the tool, the machine learning algorithms generate, on average, between zero and 32 questions with their answers. A sample of the generated questions using Lumos is shown in Figure 4.6.

Furthermore, Kumar *et al.* (2019) proposed a web-based system that generates questions that are correct in terms of meaning and grammar based on a given text, calling it ParaQG. The ParaQG system passes through four interactive stages. First, the system automatically filters out characters that cannot be used in the question-generating process, such as URLs and non-ASCII characters. In the second stage, the Stanford CoreNLP tagger tags the text's noun phrases and

[1] https://www.lumoslearning.com/llwp/free-question–answer-generator-online.html

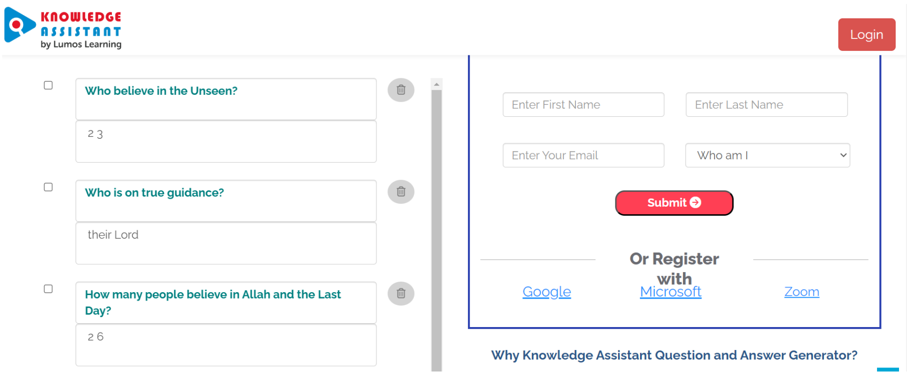



Figure 4.6: The sample of the generated questions using Lumos learning QG.

named entities. After that, the user selects the pivotal answers from these tagged sentences or any set of spans. Third, a combination of a sequence-to-sequence model, global sparse-max attention, a copy mechanism, and a dynamic dictionary is used to create questions. In the final stage, questions that do not have answers are excluded using a BERT module, while the remaining questions are grouped based on their answers, which have the same stemmed form. The system is not available at the link mentioned in the paper [1].

Furthermore, the QuestionAid tool [2] has several features. It allows users to enter a text of up to one thousand five hundred characters in twenty-four languages, such as German, Greek, Spanish, and English, which are not available in any other tools. After that, it produces five questions and their answers about the text within thirty-eight seconds. It also allows users to export the resulting questions in text format. There is only one free trial. Although users need to subscribe for a fee, there is a restriction on the number of questions that can be generated.

Additionally, PrepAI QG [3] is a platform that utilizes AI algorithms for question generation. There are no restrictions on the text entered into the tool, which can be in several formats, such as pasted text, Word documents, PDF files, and content scraped from Wikipedia. The output can include a large number of questions in different types, such as true/false, fill in the blanks, short answers, and

[1] https://www.cse.iitb.ac.in/~vishwajeet/paraqg.html
[2] https://www.question-aid.com/
[3] https://app.prepai.in/generate-questions

multiple-choice questions, with three difficulty levels: hard, medium, and easy. It provides only three free trials per day.

Furthermore, creating questions in the Quillionz QG tool [1] requires four steps. First, submit the content by pasting the text or uploading a PDF file, ensuring that the content size ranges from three hundred to three thousand words and that the text domain is specified. Then, users need to choose keywords from the text. Next, the tool marks pronouns and incomplete, long, and subjective sentences, prompting the user to replace the pronouns with nouns and modify the sentences. All these factors positively impact the quality of the questions generated. In the final step, the tool will apply ML algorithms and AI to create several questions. These can be classified into three types: multiple choice, fill-in-the-blank, and short answer questions, and users can export them as a text file. One drawback of this platform is that it requires a subscription.

Moreover, another tool that takes advantage of AI and ML algorithms is the Automatic QG [2]. The text should be pasted into the main text area. After that, words will be displayed for the user to choose keywords. Next, the user selects a set of criteria, such as the type and number of questions required. Finally, all questions will be displayed and can be exported as a Word, PDF, or text file. This tool can be used in many areas, including education and business, allowing teachers, trainers, and managers to evaluate their students or employees. It requires a subscription.

GEM Company developed a paid tool called AI-Powered QG [3] to support an automated chatbot in Japan. The motivating factor was the tendency of many companies to use chatbots in customer service, as the performance of the chatbot depends on the question–answer database. Therefore, there was a need to speed up the process of manually scanning documents to analyze and create questions and answers from them using the platform. The system's structure consists of several components, including parsing the Japanese text into sentences and then marking the sentences using a POS tagger. After that, the sentences are classified according to their importance based on the extracted features. Next, the question

[1] https://app.quillionz.com
[2] https://softwarecountry.com/our-products/automatic-question-generator/
[3] https://gemvietnam.com/case-studies/ai-question-generator-chatbot/

words are selected from the important sentences, and the sentence words are rearranged to generate a question. T5, BERT, and GPT-2 are used in this tool.

Vu & Blake (2021) designed and developed a pedagogical tool that generates questions from text to help English language learners and teach them how to construct questions correctly. The process of creating questions from text includes two stages: syntactic transformation and pronoun selection. Rules based on parse trees are used in the syntactic transformation stage. This tool generates different types of questions: tag questions, open-ended questions, and closed-ended questions. The system is implemented in Python but is not currently available.

### 4.3.2 Other Questions Type Generator Tools

Amin [1] built a QG API based on AI algorithms using RapidMiner, which specializes in creating fill-in-the-blank questions. The user places a paragraph in the text field and then presses the Test Endpoint button to automatically generate about ten blank questions with their answers in a short amount of time. The disadvantages of this QG API are that it only generates fill-in-the-blank questions in English.

Based on the above, only four freely available tools for generating short-answer questions exist. Since this is exploratory work and due to financial constraints, I have chosen to first investigate the free tools to assess their potential for building a question–answer pairs dataset for the Holy Quran in English. If these tools perform well, they will be translated into Arabic.

## 4.4 Corpus Creation

The methodology for creating QUQA and HAQA went through several stages, consisting of designing the two datasets, identifying the data sources, and collecting and cleaning the data. Figure 4.7 shows the development methodology of these two datasets, which I now go on to discuss in detail.

[1] https://rapidapi.com/darkmanaminovic/api/question-generator

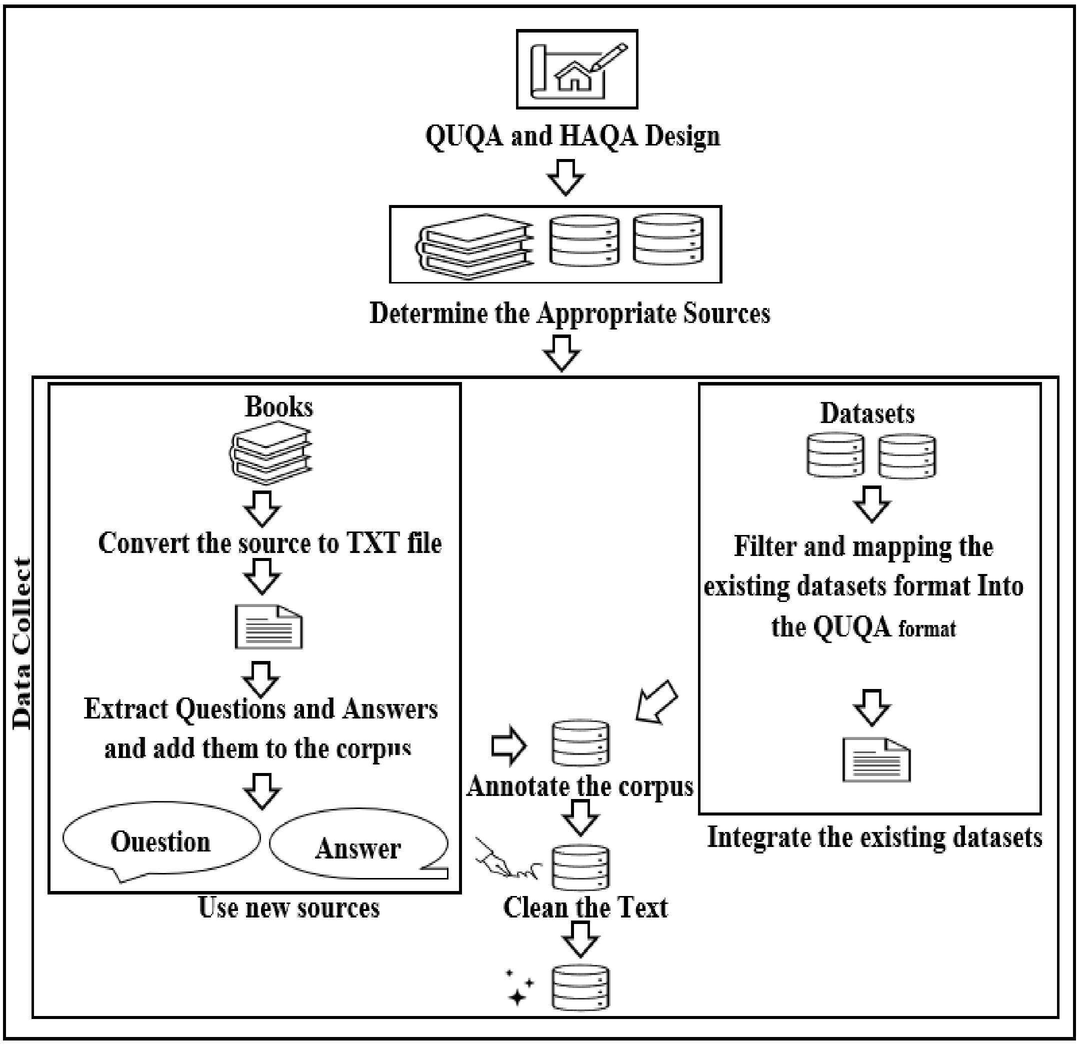



Figure 4.7: The development methodology of QUQA and HAQA.

### 4.4.1 QUQA and HAQA Design

As a starting point for building the two datasets, I must define the structure of the collection, its metadata, and the format in which it will be available. I designed the Quran dataset based on the AyaTEC and AQQAC designs, adopting the common metadata between the two corpora. Similar metadata were selected for HAQA to suit the nature of hadiths. Comma-separated values (CSV) with UTF-8 encoding were used because many systems can easily use them after converting them into XML format (Altammami, 2023a). Every record in the QUQA CSV file includes the information listed in Table 4.7 and Table 4.8, while the information in the HAQA records is shown in Table 4.9 and Table 4.10.

### 4.4.2 Identifying Data Sources

To create the corpora, I used two sources, namely books and available datasets. Many books include questions and answers about the Quran and the Hadith, but they did not meet my requirements. For example, the answers in some sources are solely in the words of an expert and do not contain evidence from the Quran or the hadiths. Additionally, I did not have permission to publish the data of some sources in my datasets. The available datasets and books matching my requirements that were used to build QUQA are as follows:

- AQQAC. This was the first Islamic dataset made available to the research community and contains answers from Quranic verses, interpretations of the verses and explanations of them in the words of an expert. This dataset file is available in XLXS and XML formats. Among the topics covered by this dataset are stories of the prophets and previous nations, Islamic legal rulings and knowledge of unseen matters (Alqahtani, 2019).

- AyaTEC. This is a specialised dataset with answers from the Holy Quran. It consists of three XML files that must be linked together. The questions relate to 11 topics, including battles, humans, stories of the prophets and faith in God (Malhas & Elsayed, 2020).

| Annotation | Description |
|---|---|
| Record_Id | The unique record number. |
| Question_Id | A question number may appear many times in this dataset due to the following:<br>1. The question has many different answers.<br>2. The question has one answer, but it is mentioned in many different verses in the Quran. |
| Question_Text | The question text |
| Quetion_Type | The type of question can be a factoid (F), a confirmation (C) or a description (D). |
| Question_Start_Word | The question keyword. |
| Answer_ID | The numbering of unique answers to the same question:<br>1. If the question has only one answer in a sense that comes totally or partially from different verses with different syntax, the numbering appears as 1.1, 1.2, 1.3 etc.<br>2. If the question has different answers in the same or different verses, the numbering appears as 1, 2, 3 etc. |
| Full_Answer | The whole answer consists of expert commentary, the Quranic verse and the hadith. |
| Expert_Commentary | An answer uses an expert's words alone. |
| Answer-Instances | The exact part of a verse that answers the question. The verse may contain more than one answer, and each answer considers an answer instance. |
| Quran_Full_Verse_ Answer | A complete verse that considers or contains the answer. |

Table 4.7: QUQA annotation part 1.

| Annotation | Description |
|---|---|
| Chapter_Name | The chapter name. |
| Chapter_Number | The chapter number. |
| Verses_Number_Start | The number of the first verse. |
| Verses_Number_End | The number of the last verse. |
| Source_Name | The name of the source. |
| Source_Link | The link to the source. |
| Credibility | Yes, if an Islamic expert has reviewed the answers; no, if they have not done so. |
| Question_ID_in_the _Original_Dataset | The question ID in the original dataset. |

Table 4.8: QUQA annotation part 2.

- 900 Questions and Answers in Managing the Verses of the Book. This is a set of questions and answers from the Quran developed by the writer due to his belief that formulating material with questions and answers increases a person's understanding of the subject (ALmuselli, 2020).

- 100 Quranic Questions and Answers. This is a set of questions and answers from the Quran developed by the writer to answer people's questions and make them more aware of their religion (Alakeel, 2018).

The books that were used to create QUQA and HAQA are as follows:

- The Doctrine of Every Muslim in a Question-and-Answer Book and the Abridged Version of the Islamic Belief from the Quran and Sunnah. This is a series of publications by Sheikh Zeno that answers the most important questions in the Muslim faith (Zeno, 2004, 2007).

- Inference on Children's Treasure. This contains a set of questions covering the following topics: the most important matters of religion, the foundations of faith, belief, the principles of jurisprudence, etiquette, dealings between people, the Prophet's biography etc (Al-Wadi, 2016).

<table>
<tr><th>Annotation</th><th>Description</th></tr>
<tr><td>Record_Id</td><td>The unique record number.</td></tr>
<tr><td>Question_Id</td><td>The question number ID may appear many times in this dataset due to the following:<br>1. The question has many different answers.<br>2. The question has one answer, but it is mentioned in many different hadiths.</td></tr>
<tr><td>Question_Text</td><td>The question text.</td></tr>
<tr><td>Question_Type</td><td>The type of question can be a factoid (F), a confirmation (C) or a description (D).</td></tr>
<tr><td>Question_Start_Word</td><td>The question keyword</td></tr>
<tr><td>Answer_ID</td><td>The numbering of unique answers to the same question:<br>1. If the question has only one answer in a sense that comes totally or partially from different hadiths with different syntax, the numbering appears as 1.1, 1.2, 1.3 etc.<br>2. If the question has different answers in the same or different hadiths, the numbering appears as 1, 2, 3 etc.</td></tr>
<tr><td>Full_Answer</td><td>The whole answer consists of expert commentary, the Quranic verse and the hadith.</td></tr>
<tr><td>Expert_Commentary</td><td>An answer uses an expert's words alone.</td></tr>
<tr><td>Hadith_Full_Answer</td><td>The entire hadith includes isnad and matn.</td></tr>
<tr><td>Hadith_Matn</td><td>The real hadith teaching.</td></tr>
<tr><td>Answer-Instances</td><td>The exact part from the hadith that answers the question. The hadith may contain more than one answer, and each answer considers an answer instance.</td></tr>
</table>

Table 4.9: HAQA annotation part 1.

| Annotation | Description |
|---|---|
| Source_Name | The name of the source. |
| Source_Link | The link to the source. |
| Credibility | Yes, if an Islamic expert has reviewed the answers; no, if they have not done so. |
| Question_ID_in_the _Orignal_Dataset | The question ID in the original dataset. |

Table 4.10: HAQA annotation part 2.

- Prayer (1770) Question and Answer. This contains people's questions related to the topic of prayer, with answers taken from the Quran and the hadiths (Al Alami, 2022).

### 4.4.3 Collecting the Data

This stage consisted of two steps, the first being to integrate the existing datasets and the second to use new sources. As mentioned earlier, there are two corpora in the Quran domain, AQQAC and AyaTEC, while the hadiths have no dataset. I wrote a Python program to convert the existing two datasets into the structure and format of my dataset. In the second step, I used the new sources to enlarge QUQA and create HAQA. The sources of the two collections consist of five books, some of which are available in text format and some not. Therefore, I wrote a Python program using OCR to convert some of the books into text format, which I reviewed manually. After that, I wrote a program that extracted questions and answers from text files and put them in my files. As a final step in this stage, I filled in the metadata using Python, or manually in some cases.

### 4.4.4 Annotating the Corpus

Metadata information was added for each record in the datasets. The metadata include the Question Type, Chapter Name, Chapter Number, Verses Number Start, Verses Number End, Source Name, Source Link in the QUQA dataset and Question Type, Source Name, Source Link, and Question ID in the original dataset. in the HAQA dataset. For more details, please see Table 4.7, Table 4.8, Table 4.9, Table 4.10.

### 4.4.5 Cleaning the Data

Cleaning data is the process of detecting and fixing errors and incorrect information. Such errors include misspellings, missing information, unwanted items, and noisy and duplicate data. This cleaning process improves the quality of the resulting data, which reflects positively on the purpose of collecting it. There are two methods for cleaning data: manual and automated. I used the manual method to discover spelling errors, missing information and duplicate information, although this approach usually takes time and effort. In addition, I used the automated approach by applying some regular expressions to remove extra spaces and non-Arabic characters. An example of a QUQA final record is shown in Table 4.11 and Table 4.12 while a HAQA example is shown in Table 4.13 and Table 4.14. I combined the answers of duplicate questions.

| Record Id | Question Id | Question Text | Question Type | Question Start Word | Answer ID | Answer Instances |
|---|---|---|---|---|---|---|
| 2350 | 1345 | ربما فرد واحد يكون سببا في إنقاذ المجتمع ونجاة أمة! آية جميلة دلت على هذا المعنى، أذكرها؟. | D | أذكر . | 1 | حتى إذا أتوا على واد النمل قالت نملة يا أيها النمل ادخلوا مساكنكم لا يحطمنكم سليمان وجنوده وهم لا يشعرون.. |
| 2350 | 1345 | Perhaps one person had succeeded in saving society and saving a nation! There is a beautiful verse that indicates this meanin. Mention it? | D | Mention | 1 | At length, when they came to a (lowly) valley of ants, one of the ants said, ‘O ye ants, get into your habitations, lest Solomon and his hosts crush you (under foot) without knowing it.’ |

Table 4.11: Example of QUQA Record– Part 1.

| **Chapter Name** | **Chapter Number** | **Verses Number Start** | **Verses Number End** | **Source Name** | **Credibility** | **Question ID in the Original Dataset** |
|---|---|---|---|---|---|---|
| النمل . | 27 | 18 | 18 | كتاب ٩٠٠ سؤال وجواب في تدبر آيات الكتاب لدريد الموصلي . | نعم . | 19.15 |
| An-Naml | 27 | 18 | 18 | 900 Questions and Answers in Managing the Verses of the Book for ALmuselli. | yes | 19.15 |

Table 4.12: Example of QUQA Record– Part 2.

| Record Id | Question Id | Question Text | Question Type | Question Start Word | Answer ID | Full Answer |
|---|---|---|---|---|---|---|
| 472 | 404 | ما هي الغزوة التي جرح فيها النبي ، وشج رأسه، وكسرت رباعيته ؟ . | F | ما . | 1 | غزوه أحد . والدليل حديث أَنَسٍ رضي الله عنه، أَنَّ رَسُولَ الله : كُسِرَتْ رَبَاعِيَتُهُ يَوْمَ أُحُدٍ، وَشُجَّ فِي رَأْسِهِ. رواه مسلم (١٧٩١). . |
| 472 | 404 | What is the name of the battle during which the Prophet, peace be upon him, took a wound to the head and had his front teeth damaged? | F | What | 1 | The Battle of Uhudl. It has been narrated on the authority of Anas that the Messenger of Allah (may peace be upon him) had his front teeth damaged on the day of the Battle of Uhudl and got a wound to his head. (Sahih Muslim, 1791). |

Table 4.13: Example of HAQA Record– Part 1.

| **Expert Commentary** | **Hadith Full Answer** | **Answer Instances** | **Source Name** | **Question ID in the Orignal Dataset** |
|---|---|---|---|---|
| غزوه أحد. | حديث أَنَسٍ رضي الله عنه، أَنَّ رَسُولَ الله : كُسِرَتْ رَبَاعِيَتُهُ يَوْمَ أُحُدٍ، وَشُجَّ فِي رَأْسِهِ. رواه مسلم (١٧٩١). | يَوْمَ أُحُدٍ | الاستدلال على كنز الأطفال للدكتور فيصل بن مسفر بن معوض الزنامي الوادعي | 595 |
| The Battle of Uhudl | It has been narrated on the authority of Anas that the Messenger of Allah (may peace be upon him) had his front teeth damaged on the day of the Battle of Uhudl and got a wound to his head. (Sahih Muslim, 1791). | On the day of the Battle of Uhudl | Inference on Children's Treasure | 595 |

Table 4.14: Example of HAQA Record– Part 2.

## 4.5 Corpus Analysis

QUQA is an Arabic question-and-answer dataset on the Holy Quran consisting of 3369 records and over 301,000 tokens. Since some questions may have more than one answer, there are 2189 questions. The answers in this corpus are extracted from 2930 verses of the Holy Quran. Accordingly, this dataset covers almost 47% of the Quran. I noticed that the questions in the new dataset are more challenging than those in the previous datasets, as shown in Table 4.11 and Table 4.12. Table 4.15 shows the comparison results between our corpus and the two existing corpora. There are 1960 single-answer and 226 multiple-answer questions. The single-answer questions are ones that have only one answer found in one or several verses in the Quran, with answers that are repeated in different places in the Quran being semantically and/or syntactically similar. The multiple-answer questions have several different answers to the question. The QUQA contains three kinds of questions: confirmation, descriptive and fact. To know the number of questions for each type, please review Table 4.16. A fact was a simple answer to a question such as where, when, etc. A descriptive question requires a more complex and detailed answer. The confirmation question, yes/no questions and boolean questions are interchangeable terms.

In addition, when I analyzed the Arabic HAQA dataset of Hadith answers, I found that there are 1598 records, over 290,000 tokens and 1366 questions (1220 single-answer and 146 multiple-answer questions). The type of questions in HAQA include: confirmation, descriptive and fact. The number of questions for each type appears in Table 4.16.

The hadiths in this collection were taken from various sources of basic hadith books; for example, there are hadiths from Al-Bukhari, Muslim, Al-Tirmidhi, Al-Nasai, Ibn Majah, Imam Ahmad, Ibn Shaybah and others.

Since the answers had been written by religious scholars, I assumed that they were correct and accurate. To enhance verification, I chose a random group of hadiths and searched for them in the LK Hadith corpus (Altammami *et al.*, 2020) and sunnah.com, and I found that they were correct.

Based on the above, the answer to the first question **RQ1** is yes, I was able to develop two datasets for the Quran and Hadith and made them available to

| Datasets | AQQAC | AyaTEC | QUQA |
|---|---|---|---|
| **# Records** | 616 | 1166 | 3369 |
| **# Questions** | 611 | 169 | 2189 |
| **#Verses in the answers** | 1232 | 1247 | 2930 |
| **% of Quran coverage** | 19% | 20% | 47% |

Table 4.15: Comparing QuQA, AQQAC and AyaTEC.

| Datasets | Confirmation | Descriptive | Fact |
|---|---|---|---|
| **QUQA** | 193 questions:<br>-143 single-answer<br>-50 multiple-answer | 1,752 questions:<br>-1618 single-answer<br>-134 multiple-answer | 241 questions:<br>-199 single-answer<br>-42 multiple-answer |
| **HAQA** | 259 questions:<br>-241 single-answer<br>-18 multiple-answer | 862 questions:<br>-748 single-answer<br>-114 multiple-answer | 245 questions:<br>-231 single-answer<br>-14 multiple-answer |

Table 4.16: The number of confirmation, descriptive and fact questions in QUQA and HAQA.

the community [1]. .

## 4.6 Corpus Enlargement

Many studies have used AQG models to create datasets, such as that of Fang *et al.* (2020), who suggested constructing open-domain question–answer pairs using the AQG approach. In this section, I analyze the effectiveness of using AQG tools to generate question–answer pairs for Quranic texts, which can be used to enlarge the dataset.

[1] https://github.com/scsaln/HAQA-and-QUQA

### 4.6.1 Dataset

This experiment was conducted on two English versions of the Quran: Yusuf Ali and Sarwar, which were obtained from Tanzil[1] as text files containing 6,236 verses. Each verse is displayed on a single line. I divided the verses into groups, and each group was placed on a separate line as input to the tool because some tools have minimum or maximum requirements for the entered text.

### 4.6.2 Experiment

The following methodology was followed:

1. Identify the free and available web service tools: the Explore AI Question Generation demo, Cathoven QG, Lumos Learning QG, and Questgen's open-source library.

2. Divide the verses into groups to fit the minimum and maximum limits of text that can be entered into each tool. The Explore AI Question Generation demo requires the input text size not to exceed 1,000 characters, while the number of words should not exceed 500 in Cathoven QG. The text should range between 50 and 500 words for Questgen QG. The minimum number of words that can be entered into Lumos Learning QG is 2,000.

3. Enter the text into the tools.

4. Export the generated questions and answers to Excel files.

### 4.6.3 Evaluation and Results

This section defines the criteria I used for the evaluation process and presents the results of this process. I used two types of standards: the first type measures the quality of the questions and answers, while the second type evaluates the overall performance of the tools.

[1]https://tanzil.net/docs/tanzil_project

1. **Measuring the Quality of the Generated Questions and Answers**:

   According to Amidei *et al.* (2018) and Zhang *et al.* (2021), the most popular human evaluation methodology is the eliciting expert judgment method. I applied this methodology by asking three annotators to rate 200 questions randomly selected from the generated questions, based on several criteria, using a rating scale from 1 to 3, where 1 is the worst score and 3 is the best. Next, I calculated an average score for each question and then an average score for all questions. For example, to calculate the score for the syntactic correctness standard of the Cathoven QG tool, each annotator read the first question generated by this tool and assigned a score between 1 and 3 for its syntactic correctness. Then I added up the scores given by the three annotators for the first question and divided the sum by three to get the average score for that question. Finally, I added all 200 average scores of the questions and divided the total by 200 to obtain the average score for the syntactic correctness standard for the Cathoven QG tool.

   The evaluation criteria I used to measure the quality of the generated questions and answers are syntactic correctness, semantic correctness, ambiguity, relevance, difficulty, and answerability. Syntactic correctness ensures that the generated questions and answers are grammatically correct, while semantic correctness ensures that the question and answer are meaningful. Ambiguity is assessed to determine if a question, when asked independently of the text, lacks ambiguity, leading to a clear answer. The relevance of the question and answer to the context is also significant. Difficulty is measured to ensure that the reader fully understands the entered text and requires some logic to answer the question. Additionally, an answerability criterion was established to verify that each answer is a plausible response to the created question. Higher scores are better for all criteria except for difficulty, where lower scores are preferred. Each criterion is evaluated independently. If the generated question contains grammatical errors but the answer is correct, it will receive a high score for the answerability criterion.

   The quality results of the questions and answers generated by the four tools are listed in Table 4.17. The Cathoven QG tool was dominant in all criteria

| Criteria | ExploreAI Question Generation demo | Cathoven QG | Lumos Learning QG | Questgen QG |
|---|---|---|---|---|
| **Syntactic Correctness** | 2.43 | **2.70** | 2.42 | 2.30 |
| **Semantic Correctness** | 2.22 | **2.44** | 2 | 1.99 |
| **Ambiguity** | 2.26 | **2.52** | 2.15 | 2.14 |
| **Relevance** | **2.98** | 2.96 | 2.86 | 2.84 |
| **Difficulty** | 1.08 | **1.20** | 1.16 | 1.07 |
| **Answerability** | 2.41 | **2.51** | 2.09 | 2.22 |

Table 4.17: Human evaluation results for questions and answers generated based on quality standards.

except the relevance, where it was outperformed by the ExploreAI QG demo by a narrow margin of 0.2%. The ExploreAI QG demo achieved the second-highest score in most criteria, except the difficulty criterion. It was beaten by the Lumos Learning QG, which received the third-highest score in the rest of the standards. The Questgen QG results are at the bottom of the list.

2. **Measuring the Performance of Tools:**

   I also measured the tools' performance using several criteria, including the variety of question types, usability, the total number of questions, and the actual duration if assuming the tools are running continuously. First, the variety of generated question types is an indication of a tool's capability. Usability refers to how easily a user can use a tool effectively and efficiently. The total number of questions generated for the entire Quran is also an important criterion because it indicates performance. The last metric refers to the time it takes to create questions for the whole Quran.

Table 4.18 presents the results of the performance evaluation of the four tools based on experiments conducted to generate question–answer pairs for the Holy Quran. The top tools are Cathoven QG and Lumos Learning QG, based on two measures. Cathoven QG is the best tool in terms of the diversity of question types and ease of use. In comparison, Lumos Learning QG excels in the total number of questions generated and duration. It is followed by the ExploreAI QG demo and Questgen QG, which perform better on other criteria. ExploreAI QG demo is superior in ease of use, while Questgen QG excels in duration. On average, Lumos Learning QG and Cathoven QG outperformed the ExploreAI QG demo and Questgen QG. The difference in superiority is only in terms of one criterion.

To summarize the above, on average, Cathoven QG is the best in terms of tool performance and the quality of the generated text. Regarding the other tools, ExploreAI QG demo and Lumos Learning QG are at the same level, followed by Questgen QG.

### 4.6.4 Discussion and Analysis

This section analyzes the results obtained by the four web service tools and discusses the possible reasons for these results.

As previously stated, Cathoven QG achieved the highest results compared to the other tools in terms of the quality of the generated text. For example, when I entered a number of verses, I found that each tool generated different question–answer pairs, as shown in Figure 4.8. Cathoven QG created a question–answer pair that is correct in terms of meaning and syntax, complete in terms of relevance to the text, and clear in terms of understanding; there is no ambiguity when asking the question independently. In comparison, the other tools suffered from problems: first, there was incorrect sentence structure or meaning, for example, with Lumos Learning QG and Questgen QG. Second, the ambiguity was sometimes due to the presence of pronouns in the question or answer, as seen with the ExploreAI QG demo. Finally, there was occasionally a wrong answer, such as with Lumos Learning QG and Questgen QG.

| Criteria | ExploreAI Question Generation demo | Cathoven QG | Lumos learning QG | Questgen QG |
|---|---|---|---|---|
| **Variety of Question Types** | Who, What, How, Where, Whose, Why, When, Which, Whoever | **Who, What, How, Where, Whose, Why, When, Which, Whoever, Whom** | Who, What, How, Where, Whose, Why, When, Which, Whoever | Who, What, How, Where, Whose, Why, When |
| **Usability** | **It is easy to use** | **It is easy to use** | The user needs time to learn how to use the tool. | A user with a background in Python programming can handle the tool and complete the task successfully and efficiently. |
| **Total Number of Questions** | Sarwar: 2,761 Yusuf Ali: 4,174 Total: 6,935 | Sarwar: 6,815 Yusuf Ali: 5,902 Total: 12,717 | **Sarwar: 8,537 Yusuf Ali: 5,062 Total: 13,599** | Sarwar ; 3,675 Yusuf Ali: 3,659 Total: 7,334 |
| **Duration** | Five days | Two days | **12 hours** | **12 hours** |

Table 4.18: Human evaluation results for the tools' performance.

**Context from the version of Yusuf Ali:** 2|31|And He taught Adam the names of all things; then He placed them before the angels, and said: "Tell me the names of these if ye are right." 2|32|They said: "Glory to Thee, of knowledge We have none, save what Thou Hast taught us: In truth it is Thou Who art perfect in knowledge and wisdom." 2|33|He said: "O Adam! Tell them their names." When he had told them, Allah said: "Did I not tell you that I know the secrets of heaven and earth, and I know what ye reveal and what ye conceal?"

**Al-Tabari's interpretation (Al-Tabari, 1954):** God taught Adam the names of everything, such as the sea and the mountain. Then God presented the owners of the names to the angels and asked them to tell him their names, but the angels could not answer because they do not know the unseen. Then God asked Adam to tell them the names, so he told them.

| **Web Service Name** | **Answer** |
|---|---|
| Cathoven QG | Who taught Adam the names of all things?<br>Allah |
| ExploreAI Question Generation demo | What did Adam tell them?<br>their names |
| Lumos learning QG | Who did He teach the names of all things??<br>Adam |
| Questgen QG | Who did He teach Adam the names of all things?<br>angels |

Figure 4.8: A sample of generated questions using the Cathoven QG, the ExploreAI QG demo, the Lumos Learning QG, and the Questgen QG for the same verses.

All the tools generate texts that suffer from issues affecting their performance. The difference between them lies in the types of problems and the amount of text that gives rise to these issues, such as the following examples:

1. Sometimes, the Cathoven QG generates incomplete answers. For example, it generated the following question and answer from the script below:

   **Question**: 'What did Satan do to the children of Israel?'

   **Answer**: 'make them slip from the (garden) and get them out of the'

   **Script**: 'Then did Satan make them slip from the (garden), and get them out of the state (of felicity) in which they had been. We said: "Get ye down, all (ye people), with enmity between yourselves. On earth will be your dwelling-place and your means of livelihood – for a time."' (Verses 36 of Surat Al-Baqarah in the version of Yusuf Ali)

   However, **the complete, correct answer** is 'make them slip from the (garden) and get them out of the state (of felicity)'.

2. One of the problems I encountered was incorrect answers. For example, the ExploreAI QG demo tool generated the following question–answer:

   **Question**: 'Who is an enemy to Allah and His angels and messengers?'

   **Answer**: 'Gabriel and Michael.'

   **Script**: 'Whoever is an enemy to Allah and His angels and messengers, to Gabriel and Michael, Lo! Allah is an enemy to those who reject Faith.' (Verse 98 of Surat Al-Baqarah in the version of Yusuf Ali).

   As we can see, the answer is wrong. **The correct answer** is 'those who reject Faith.' This tool may not be able to understand and analyze bracketing commas.

3. The Lumos Learning QG generated the following question–answer:

   **Question**: 'Who will not grasp the Message?'

   **Answer**: 'Men of understanding.'

**Script**: 'He granteth wisdom to whom He pleaseth; and he to whom wisdom is granted receiveth indeed a benefit overflowing; but none will grasp the Message but men of understanding.' (Verse 269 of Surat Al-Baqarah in the version of Yusuf Ali)

As we can see, the question is wrong, and **the correct question** is: 'Who will grasp the Message?' We must conclude that the Lumos Learning QG could not correctly analyze the negation and affirmation texts.

4. Sometimes the answer appears in the question, and the grammatical structure of the question is incorrect, as in the following question–answer generated by the Questgen QG:

   **Question**: 'Who did Adam learn from his Lord's words of inspiration?'

   **Answer**: 'Adam.'

   **Script**: 'Then learnt Adam from his Lord words of inspiration, and his Lord turned towards him; for He is Oft-Returning, Most Merciful.' (Verse 37 of Surat Al-Baqarah in the version of Yusuf Ali)

Regarding the results of the four web services and their performance in general, as presented in Table 4.16, I can note several points:

- I noticed that all the tools generated the same types of questions: who, what, how, where, whose, why, and when, while the ExploreAI QG demo and Lumos Learning QG tools can also generate questions that start with 'whoever' and 'which.' In addition to all of the above, the Cathoven QG tool can generate questions that begin with 'whom'.
- The Cathoven QG and ExploreAI QG demo were easy to use. Users only need to copy and paste text and click the button, but Lumos Learning QG requires registration and time to access the tool. Questgen QG requires users to have programming skills in Python to use the library.
- The Cathoven QG and Lumos Learning QG tool generated a similar number of questions—around 13,000—while Questgen QG and ExploreAI QG demo produced half that number.

- Lumos Learning QG and Questgen QG were the fastest tools to run.

The disadvantages of these tools are that the resulting questions and answers are not of high quality, contain many errors, and are too straightforward. Therefore, I did not add these questions to my collection.

## 4.7 Conclusion

This chapter surveys the existing Islamic question–answer pairs datasets and compares them using thirteen measures. The following points summarize our findings: (1) To the best of our knowledge, no current public question–answer pairs corpus specialized in Hadith is available. The existing corpora cover only Arabic Quran questions. (2) Hundreds is the average number of questions, which is considered very small. (3) Available datasets cover only a small portion of the Quran. There is a need for a Hadith-specific dataset, as well as a larger dataset that encompasses a broader portion of the Qur'an and contains a greater number of questions.

Therefore, two datasets were developed to overcome the limitations of existing corpora. QUQA and HAQA are two datasets that contain questions and answers about the Holy Quran and Hadith, respectively. Since these corpora include more than 4,900 records, they are considered to be the largest Islamic corpora available to the research community. These two collections contain more than 590,000 tokens.

An attempt has been made to expand the Holy Quran dataset by using four AQG tools, namely the ExploreAI QG demo, Cathoven QG, Lumos Learning QG, and Questgen QG. However, the question–answer pairs generated were not of high quality.

# Chapter 5

# Evaluating the Pre-trained Language Models (PLMs) for Passage Retrieval Task

## 5.1 Introduction

Passage ranking is based on retrieving a passage from a larger corpus that contains an answer to a specific question. This task has not been fully explored in the CA language. Due to the small size of the only available dataset in CA, I created two CA datasets. These datasets were used to train eight Arabic pre-trained transformer-based models, which were used in different approaches to passage ranking to study their performance. The approaches were the dense representation approach, the relevance classification approach, and the hybrid approach. Finally, I studied the effect of the optimization methods: the transfer learning or ensemble technique. I observed some substantial improvements in the results, but I did not conduct strict statistical significance testing due to the small sample size. This chapter was built on the following publication:

Alnefaie S, Atwell E, Alsalka MA. Qur'an Passage Ranking Using Transformer Models **Accepted in** The Eighth International Conference on Arabic Language Processing, (ICALP 2024).

## 5.2 Background

In the PR task, the system has two inputs: a question and a set of passages. The output of the system is a ranked list of passages that answer the question. PR is based on retrieving an ordered list of passages that answer a question from a larger corpus. This list of passages is sorted based on a score assigned to each. Each score is computed based on a chosen approach (Alsubhi *et al.*, 2022). The PR model architecture is shown in Figure 5.1.

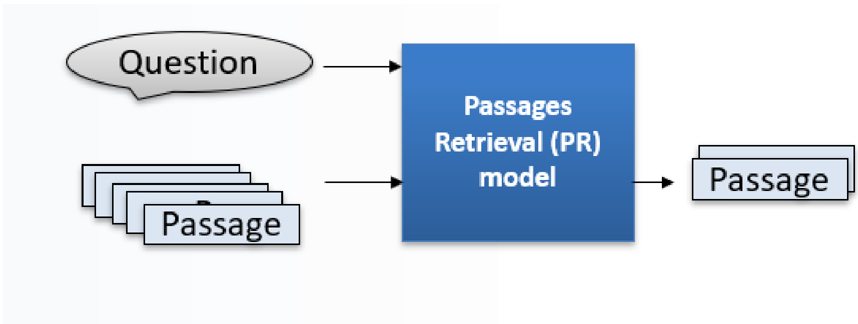



Figure 5.1: Passage retrieval model architecture.

## 5.3 Related Work

There are three general categories of approaches: exact match, semantic match, and hybrid approach.

### 5.3.1 Exact match Approach

Exact match is a traditional approach based on statistical information about the query terms and corpus or set of passages; examples include Best Match 25 (BM25) (Robertson *et al.*, 2009) and term frequency-inverse document frequency (TF-IDF) (Salton & Buckley, 1988). This approach suffers from a vocabulary mismatch problem, which means it is unable to retrieve passages or documents that contain different words that are semantically similar to those in the question.

To tackle this challenge, researchers use several techniques, including query expansion, passage expansion, and semantic match. The idea of expanding the query or passage is based on adding synonyms of terms found in a query or passage. Recently, the semantic match approach has shown promising results when used with pre-trained transformer-based models to capture meaning (Yates *et al.*, 2021).

### 5.3.2 Semantic Match Approach

The PR semantic match method is an approach based on dense representations and relevance classification (Yates *et al.*, 2021). These two concepts are described in the following paragraphs.

- **The Dense Passage Retrieval Approach:**

  The Dense Passage Retrieval (DPR) approach involves three steps: first, the deep learning model converts the question and all passages into numerical representations or vectors. Second, it measures the similarity between the question vector and each passage. Finally, it selects the top similar passages. The architecture of the DPR approach is shown in Figure 5.2. The embedding models in the DPR could be BERT variants or OpenAI models. The BERT pre-trained model used in this approach is called a bi-encoder model (Karpukhin *et al.*, 2020).

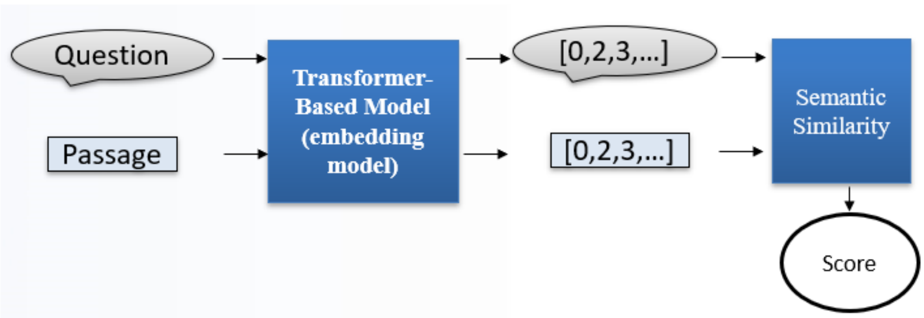



Figure 5.2: The architecture of the Dense Passage Retrieval approach

Karpukhin *et al.* (2020) studied the performance of the DPR technique using BERT as a bi-encoder model with the following English datasets: Stanford Question Answering Dataset (SQuAD), CuratedTREC, WebQuestions, TriviaQA, and Natural Questions. The results showed that the DPR approach outperformed BM25 in terms of the accuracy of top-20 passage retrieval by 9%-19%.

Additionally, Alsubhi *et al.* (2022) proposed applying the DPR approach to MSA. They used the dpr-ctx_encoder-bert-base-multilingual BERT model and utilized the Arabic Reading Comprehension Dataset (ARCD) and the Arabic Read Typologically Diverse Question Answering – Gold Passage (TyDiQA-GoldP) dataset. They conducted many experiments using different datasets for fine-tuning and testing. In fine-tuning, they used the ARCD alone, the TyDiQA-GoldP dataset alone, and both ARCD and TyDiQA-GoldP together. For testing, they used the ARCD test set, the ARCD development set, the TyDiQA-GoldP test set, and the TyDiQA-GoldP development set, using top-20 and top-100 PR accuracy as evaluation metrics. The results from all these experiments showed that the DPR approach outperformed TF-IDF.

Zekiye & Amroush (2023) used the 'text-embedding-ada-002' model, OpenAI's embedding model. The TREC AyaTEC development set was used as the test dataset. It achieved a mean reciprocal rank (MRR@10) score of 0.267 and a MAP@10 score of 0.109, which was worse than BM25. The BM25 model achieved 0.170 MAP and 0.313 MRR scores. Mahmoudi *et al.* (2023) recommended using AraBERT as a bi-encoder model. They fine-tuned the model with the Arabic portion of Mr. TyDi's dataset and the training set of the TREC AyaTEC, a CA dataset. The test dataset used in this experiment was the TREC AyaTEC development set, which is a Quran dataset. The result of this model was a 0.145 MAP@10 and 0.257 MRR@10, which was worse than BM25. The TREC AyaTEC has a limitation in the number of questions, with only 199 questions. There is a relationship between the size of the training data and the performance of pre-trained transformer-based models; the larger the dataset, the better the model's performance (Malhas *et al.*, 2022). Therefore, I decided to build a larger Quran CA dataset for training.

- **The Relevance Classification Approach:**

The pre-trained model in the relevance classification approach is used as a classifier and is called a cross-encoder model. The input to the model is the question and passage, and the output is the semantic score of these two inputs. The architecture of the relevance classification approach is shown in Figure 5.3.

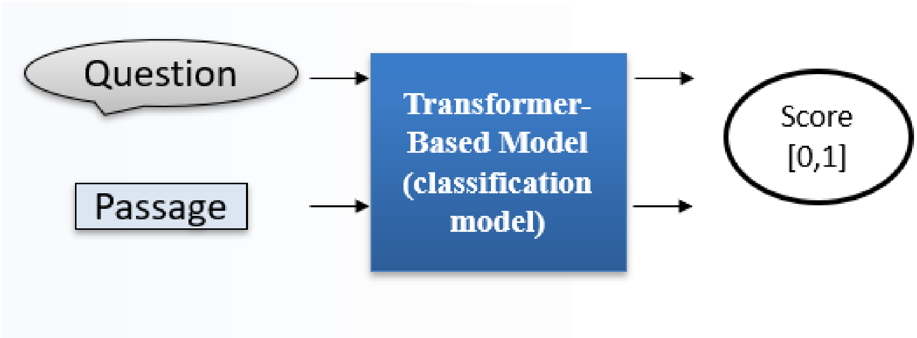



Figure 5.3: The architecture of the relevance classification approach

Grundmann *et al.* (2021) investigated the performance of the BioBERT transformer-based model when used as a cross-encoder model and a bi-encoder model for the clinical domain. In general, the recall of the top-ranked passage (R@1) of the cross-encoder model outperformed the bi-encoder model, as well as the classical approaches TF-IDF and BM25, in the WikiSection, Medical Question Answering Dataset (MedQuAD), HealthQA, and Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) datasets. The highest score achieved by the cross-encoder model was 77.90 R@1 with the MedQuAD dataset, while the bi-encoder model achieved 46.74 R@1, BM25 achieved 31.98 R@1, and TF-IDF achieved 24.87 R@1.

Research on the DPR approach, which achieved good results compared to the exact match approach (BM25 or TF-IDF), has been conducted in a small number of languages, such as English (Karpukhin *et al.*, 2020) and MSA (Alsubhi *et al.*, 2022). In CA, research concluded that the results of this approach were poor, even worse than BM25, when the TREC AyaTEC Quran dataset was used for testing and training (Mahmoudi *et al.*, 2023; Zekiye & Amroush, 2023). Research on the relevance classification approach has been conducted only in English Grundmann *et al.* (2021). To my knowledge, no study has verified the performance of these approaches in answering questions about the Hadith. Thus, there are still several unanswered questions in applying these approaches in CA. The first questions addressed in this study were the following:

- **RQ5.1**: Does training a bi-encoder model in dense representation or a cross-encoder model in relevance classification on a larger Quranic dataset improve the results to become better than the traditional statistical method BM25 for the PR task?
- **RQ5.2**: Is using the dense representation approach or the relevance classification approach better than the traditional statistical approach BM25 in CA for the PR task of Hadith?

These questions are numbered starting from 5 because they are in Chapter 5 and are not related to the general questions.

### 5.3.3 Hybrid Approach

This approach is sometimes called the Multi-stage Approach (Yates *et al.*, 2021). The hybrid approach is based on the retrieve-and-re-rank architecture, which consists of two stages. First, the top passages that answer the question are retrieved from the corpus using an initial approach, such as the traditional BM25 or DPR. These top candidate passages are then re-ranked using a second approach, such as relevance classification or DPR approach (Yates *et al.*, 2021). Alnefaie *et al.* (2023a) suggested using the DPR approach and comparing it with the hybrid approach. In the hybrid approach, they used ArabicBERT, CAMeL-BERT, AraBERT, or CL-AraBERT as a bi-encoder model to retrieve the relevant passages and re-ranked them using "mmarco-mMiniLMv2-L12-H384-v1" as a cross-encoder model. They trained the models using the training set from the TREC AyaTEC dataset and tested the model using the test set from the same dataset. The results showed that the DPR method was better than the hybrid approach in all four experiments with the Arabic models, with the best model being CL-AraBERT, which achieved 0.124 MAP and 0.375 MRR when used as a bi-encoder only.

Nogueira & Cho (2019) and Xie *et al.* (2023) proposed a PR model that retrieved the top passages using BM25, which were then re-ranked with a BERT model as a cross-encoder. The first study was conducted in English, while the second was in Chinese. The model proposed by Nogueira & Cho (2019) outperformed BM25 in terms of mean reciprocal rank (MRR@10) and MAP on the TREC-Complex Answer Retrieval (TREC-CAR) dataset and the MSMARCO (Microsoft Machine Reading Comprehension) dataset. The performance of the proposed model could potentially double that of BM25. For example, with the TREC-CAR dataset, the proposed model produced a MAP score of 33.5, while BM25 generated a MAP score of 15.3. Furthermore, the Chinese hybrid approach outperformed BM25 by 0.15 MRR@10 on the T2Ranking dataset.

In the experiments conducted in the CA language, the hybrid approach that used a bi-encoder model to retrieve the relevant passages and re-ranked them using a cross-encoder model yielded worse results than using the bi-encoder model alone (Alnefaie *et al.*, 2023a). In the experiments involving the Chinese and English languages, the hybrid approach retrieved the top passages using BM25, which were then re-ranked with a cross-encoder model. This hybrid approach outperformed BM25 (Nogueira & Cho, 2019; Xie *et al.*, 2023). Thus, the third research question was:

- **RQ5.3**: Does the hybrid approach (retrieved the top passages using BM25, which were then re-ranked with a bi-encoder model or a cross-encoder model) outperform BM25 in CA for the PR task?

The transfer learning technique and the ensemble technique were used to improve the performance of pre-trained transformer-based models in many NLP tasks, such as PR. The idea behind the transfer learning technique is to use additional datasets that are larger than the basic dataset of the experiment. When models are trained on large datasets, their performance improves (Malhas, 2023). Alnefaie *et al.* (2023a) proposed using the ensemble technique on the results of Arabic language models when they were used as bi-encoders. The proposal achieved the highest results when they combined the results of ArabicBERT and

CL-AraBERT, applying the ensemble technique to achieve 0.534 MAP and 0.73 MRR. They tested the model using the development set from the TREC AyaTEC dataset. Thus, the fourth and fifth research questions were as follows:

- **RQ5.4**: Does the transfer learning technique, using my new and larger Quran dataset, and/or the ensemble technique improve the performance of pre-trained transformer-based models for the PR task of the Holy Quran compared to their performance without these techniques?
- **RQ5.5**: Does the transfer learning technique and/or the ensemble technique improve the performance of the pre-trained transformer-based models for the PR task of the Hadith?

## 5.4 The Proposed Model

I conducted two separate experiments. The first experiment focused on finding the related passages to the question from the Holy Quran, while the second experiment focused on finding the related passages from the Hadith. The methodology used in this proposal is based on employing the following three approaches and comparing their results:

1. **DPR Approach:** First, dense representations are utilized in the DPR technique. The concept behind DPR is to generate a representation vector for the query and all passages in the corpus using the transformer model. To compute the similarity score for each passage with a query, the dot product of the passage and query vectors is then calculated. Finally, the passages are sorted based on these scores.

   In this proposed model, I used eight Arabic pre-trained transformer-based models. The models were AraBERT Base (Antoun *et al.*, 2020) [1], ArabicBERT (Safaya *et al.*, 2020) [2], AraELECTRA (the Arabic version of efficiently learning an encoder that classifies token replacements accurately

[1] https://huggingface.co/aubmindlab/bert-base-arabert
[2] https://huggingface.co/asafaya/bert-base-arabic

(ELECTRA)) (Antoun *et al.*, 2021) [1], QCRI Arabic and Dialectical BERT (QARiB) (Abdelali *et al.*, 2021) [2], CAMeL-BERT (Computational Approaches to Modelling Language - Bidirectional Encoder Representations from Transformers) (Inoue *et al.*, 2021) [3], Arabic BERT (ARBERT), Masked Arabic BERT (MARBERT) (Abdul-Mageed *et al.*, 2021) [4], and CL-AraBERT (Malhas & Elsayed, 2022) [5]. In the experiment, the parameters were set as follows: the batch size was 16, the learning rate was 1e-4, and the number of epochs was 4.

The process of building a training dataset consisted of two steps. First, I extracted all the question-passage pairs from the training dataset. Second, I formatted each record as follows: [question, positive passage +, negative passage -]. The question and positive passage (relevant to the question) were extracted from the question-passage pairs in the dataset, while the negative passage (not relevant to the question) was chosen randomly.

For example, if the golden dataset was:
question 1 passage 1,
question 2 passage 2, ...
The new record in the training step would be:
[question 1, passage 1 +, passage 2 -]

I then used this dataset to fine-tune each of the eight transformer-based models used in this experiment. After that, I built an index for all Quran or Hadith passages. The file that contains all the Quran passages is the Quranic Passage Collection (QPC) file from TREC AyaTEC [6]. The file that contains all the Hadith passages is the Hadith Passage Collection (HPC) file, described in Section 5.7.1. I used the model to map each passage to a d-dimensional, real-valued vector and stored these vectors. At runtime,

[1] https://huggingface.co/aubmindlab/araelectra-base-generator
[2] https://huggingface.co/ahmedabdelali/bert-base-qarib
[3] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca
[4] https://huggingface.co/collections/UBC-NLP/arbert-and-marbert-6615a218e0f403ced3894ba0
[5] https://huggingface.co/qahq/CL-AraBERTv0.1-base
[6] https://gitlab.com/bigirqu/quran-qa-2023/-/blob/main/Task-A/data/Thematic_QPC/QQA23_TaskA_QPC_v1.1.tsv?ref_type=heads

a d-dimensional vector was generated for each question. To determine the closest passage vectors to the question vector, I calculated the dot product of their vectors. The passages were then sorted based on the calculated scores. Finally, I retrieved the passage with the highest score. The steps of the DPR approach are shown in Figure 5.4.

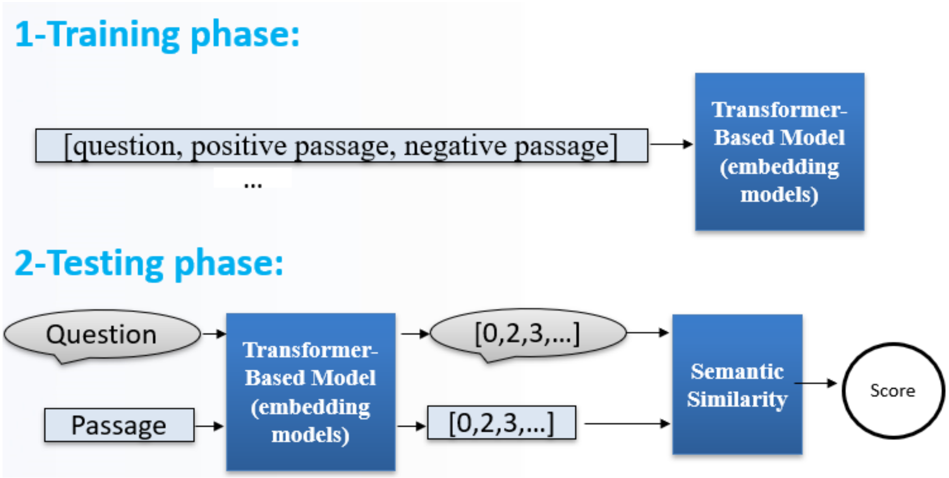



Figure 5.4: The steps of the Dense Passage Retrieval approach

2. **Relevance Classification Approach:** The approach is based on relevance classification. The pre-trained transformer-based model in this approach is used as a classifier and is referred to as a cross-encoder model. The input to the model consists of two sentences, and the output from the model is a score ranging from zero to one, indicating the semantic similarity between them. In the PR task, each passage in the dataset is entered into the model along with the question. The passages are then sorted based on the scores obtained when they are retrieved as relevant passages for each question (Yates *et al.*, 2021).

   In the relevance classification approach, I used the eight previously Arabic pre-trained transformer-based models as cross-encoder models. In addition, I used six multi-language transformer-based models: mmarco-mMiniLMv2-L12-H384-v1 [1], stsb-roberta-base (robustly optimised BERT-pretraining

[1] https://huggingface.co/corrius/cross-encoder-mmarco-mMiniLMv2-L12-H384-v1

approach) [1], ms-marco-MiniLM-L-12-v2 [2], sts-bdistilroberta-base [3], qnli-distilroberta-base [4], and ms-marco-TinyBERT-L-2-v2 [5].

First, each record in the training dataset was represented as follows: [question, passage, 1]. The question and passage pairs were extracted from the training datasets. A value of 1 indicated that they were semantically similar. To build the negative samples of non-relevant pairs, I paired each question with a passage that answered another question in the datasets, with the following format: [question, passage, 0]. I then fine-tuned the transformer-based model using this dataset. After that, I built an index for all passages in the Holy Quran or in the or Hadith. At runtime, the model computes the semantic similarity scores for the questions with each passage. These scores are then used to sort the passages. Finally, the top passages are extracted. The steps of the relevance classification approach are shown in Figure 5.5.

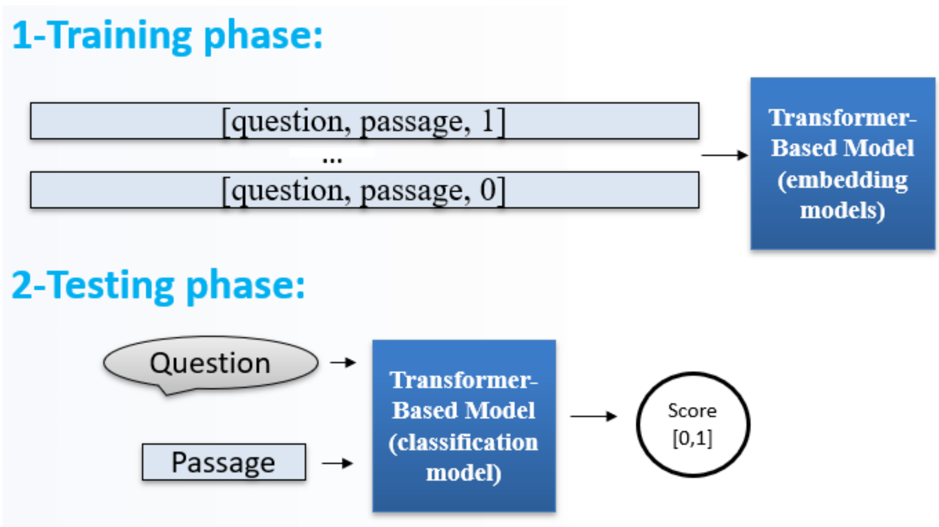



Figure 5.5: The steps of the relevance classification approach

[1]https://huggingface.co/cross-encoder/stsb-roberta-base
[2]https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
[3]https://huggingface.co/cross-encoder/stsb-distilroberta-base
[4]https://huggingface.co/cross-encoder/qnli-distilroberta-base
[5]https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L-2-v2

3. **Hybrid Approach:** First, I used BM25 to retrieve the passages relevant to the question. The top relevant passages were then selected. Finally, I implemented the previously described DPR or relevance classification approach to re-rank the passages. In fact, the number of top relevant passages selected is a factor that affects performance. Therefore, I ran experiments to determine how adjusting the number of selected documents improved the results with this dataset. The steps of the hybrid approach are shown in Figure 5.6.

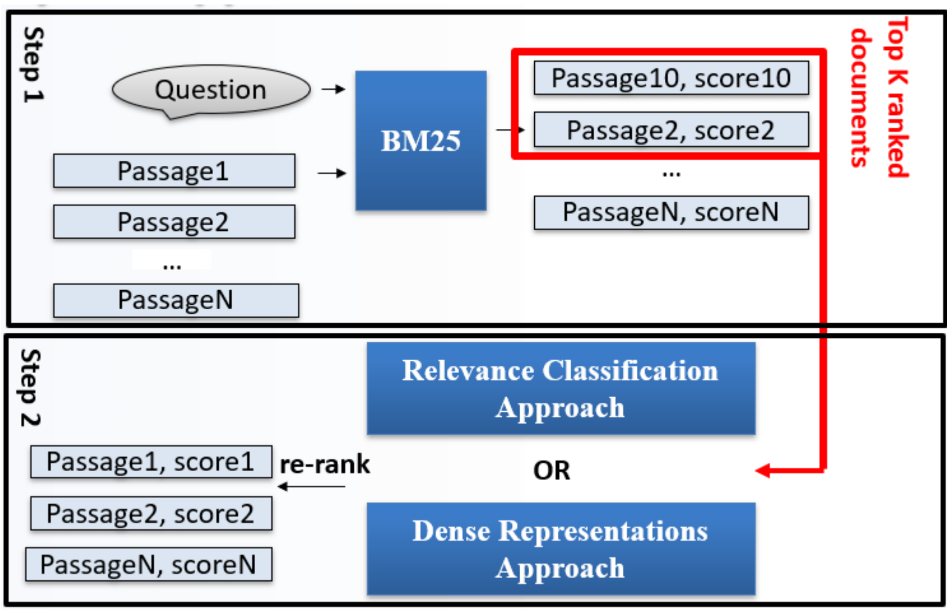



Figure 5.6: The steps of the hybrid approach

**Improvement Techniques:** There are two possible techniques to improve the results: the transfer learning technique and the ensemble technique.

- **Transfer Learning Technique:** The idea behind the transfer learning technique is that if the dataset available for training is small, the model can be trained on a larger dataset, the weights of the model are saved, and it can then be fine-tuned on the smaller basic dataset (Malhas & Elsayed, 2022). In this study, I applied this technique by building four models. The only difference between these models was the training dataset:

  - **Experiment 1:** In the first model, I trained the AraBERT Base model to be used as a bi-encoder in the hybrid approach using only the primary dataset (TREC AyaTEC), which is considered a CA dataset. In the second model, I trained the same model with both TREC AyaTEC and QuranTrec, which are considered CA datasets. The third model was fine-tuned using only the MSA datasets ARCD and Arabic SQuAD. The two CA datasets and the two MSA datasets were used together to fine-tune the fourth model.
  - **Experiment 2:** In the initial model, I trained the AraBERT base model to function as a bi-encoder within a hybrid framework, utilising the primary dataset HadithTrec, which is classified as a CA dataset. For the second model, I trained the AraBERT model on both the HadithTrec and QuranTrec datasets, both of which are also categorised as CA datasets. The two MSA ARCD and Arabic SQuAD datasets were used to fine-tune the third model. Finally, the fourth model was fine-tuned using a combination of the two CA datasets and MSA datasets.

- **Ensemble Technique:** To implement this technique, I followed these steps: First, for each question, I selected the top 20 candidate passages from each individual model. Next, for each of those passages, I computed the summation of its score from all the models. I then calculated the average score for each passage. Finally, the top 10 candidate passages were considered the retrieved passages for the question (Alnefaie *et al.*, 2023a; ElKomy & Sarhan, 2022).

**- Experiment 1:** The TREC AyaTEC is the testing dataset. The TREC AyaTEC, QuranTrec, SQuAD and ARCD are the fine-tuning datasets.
**- Experiment 2:** The HadithTrec is the testing dataset. The HadithTrec, QuranTrec, SQuAD, and ARCD are the fine-tuning datasets.

## 5.5 Evaluation

I used MAP and MRR as the evaluation metrics for these experiments because all the research in the field of the Qur'an has utilized them. I considered MAP to be the main metric (Liu *et al.*, 2009; Malhas *et al.*, 2023).

- **Mean Reciprocal Rank (MRR):** For query $q$, $MRR$ is a measure that is concerned with the rank of the highest-ranked related document in the list as shown in Eq 5.1.

$$MRR(q) = \frac{1}{rank} \tag{5.1}$$

  $MRR$ for $n$ queries as shown in Eq 5.2.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \tag{5.2}$$

  where $rank_i$ refers to the rank of the relevant document for the $i$ -th query.

  **Mean Reciprocal Rank at $k$ ($MRR@k$):** This metric is concerned with the rank of the highest-ranked relevant document in the first $k$ documents in the retrieve list. In this chapter's experiments, a metric $MRR@10$ was used.

- **Mean Average Precision (MAP):** For query $q$, $MAP$ is a measure assesses whether all relevant items are generally ranked at a high position as shown in the following:

  **To Compute MAP**:

  1. **For query $q$:**
     - **Precision at position $k$ ($P@k$):**

$$P@k(q) = \frac{\#\{\text{relevant documents in the top k positions}\}}{k} \tag{5.3}$$

- **The Average Precision *AP*:**

$$AP(q) = \frac{\sum_{k=1}^{m} P@k(q) \times l_k}{\#\{\text{relevant documents}\}} \tag{5.4}$$

Where $m$ denotes the total number of passages or documents relevant to the query $q$, while $l_k$ reflects a binary evaluation of the relevance of the passage or document at position $k$.

2. **For $n$ queries :** MAP is the mean value of AP overall questions.

In this chapter's experiments, a metric $MAP@10$ was used.

## 5.6 Experiment 1: Answering Questions from the Holy Quran

### 5.6.1 Datasets

In this research, four Arabic datasets were used in the experiments. The first dataset was used for training and evaluation, while the remaining three datasets were used only for training. The questions in the first two datasets were written in MSA, while the passages were written in CA. The language in the last two datasets was MSA. Dataset details are as follows:

1. **TREC AyaTEC:** This dataset consists of three files: the Quranic passage collection (QPC), AyaTEC questions, and query relevance judgments (QRels). The QPC comprises 1,266 passages that represent the entire Holy Quran. The Quran was divided into these passages using the Thematic Holy Quran[1], which organizes the Quran verses into groups according to the topics of the verses. Each passage has an ID number. The question file includes 199 questions, each with its own ID. The questions were divided into 70% for training, 10% for development, and 20% for the test sets. The QRels file is composed of 1,132 gold records, with each record being a pairing of a question ID and the ID of the passage that answers it (Malhas &

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

Elsayed, 2020; Malhas, 2023). This dataset is the only one available for PR tasks in CA, but it has a limitation in the number of questions, as there are only 199 questions. There is a relationship between the size of the training data and the performance of pre-trained transformer-based models; the larger the dataset, the better the model's performance (Malhas *et al.*, 2022).

2. **QuranTrec:** It is an extended version of the QUQA dataset. The QUQA is a question-answer pairs dataset, while QuranTrec consists of question-passage-answer triplets regarding the Arabic Holy Quran. The QuranTrec is a PR dataset. The process of finding the passage for each answer in the QUQA to fit the QuranTrec was as follows:

    (a) Determined the beginning and end numbers of the verses in the answer.

    (b) Based on the number of the first and last verses in the answer, I select the appropriate passage from the QPC file of the TREC AyaTEC dataset. The appropriate passage is the one that includes all the verses mentioned in the answer. Sometimes, the first part of the answer was located in one passage and the second part in the next passage. In this case, my passage containing the answer consisted of the first passage and part of the second passage.

    (c) Added the appropriate passage for each record in the dataset.

    Once this was done for all records, they were added to the QuranTrec dataset. Only the question and the passage pairs were extracted from the dataset.

3. **ARCD:** This dataset consists of 1,395 question-passage-answer triplets for Wikipedia passages (Mozannar *et al.*, 2019). Only the question and passage pairs were extracted from the dataset.

4. **Arabic SQuAD:** This dataset consists of 48.3k question-passage-answer triplets for the machine translation of the SQuAD dataset (Mozannar *et al.*, 2019). Only the question and passage pairs were extracted from the dataset.

### 5.6.2 Results

To answer **RQ5.1**, which relates to comparing the performance of the DPR and relevance classification approaches with BM25 when the models were further trained using the QuranTrec dataset, I conducted three experiments. First, I built a PR model using BM25. Second, I built a PR model using the DPR approach; those results appear in the "Bi-Encoder" column in Table 5.1. Third, I used the relevance classification approach to build a PR model, and the results of using Arabic models appear in the "Cross-Encoder" column in Table 5.1, while the results of using multilingual models appear in Table 5.2. BM25 obtained 0.170 MAP@10 and 0.313 MRR@10.

From the results, I noted that five out of eight models achieved higher MAP@10 results than BM25 when used in the DPR approach (results are colored in blue). The models were AraBERT Base, ArabicBERT, CAMeL-BERT, AraELECTRA, and CL-AraBERT. The highest-performing model among them was AraBERT, which achieved 0.244 MAP@10 and 0.413 MRR@10. From Table 5.1, I observed that all the Arabic models achieved worse results than BM25 when used with the relevance classification approach. The best of them was CAMeL-BERT, with 0.125 MAP@10 and 0.219 MRR@10. The only model that outperformed BM25 among the multilingual models was qnli-distilroberta-base, with 0.198 MAP@10, as shown in Table 5.2.

From the above, the answer to question **RQ5.1** is that, when using the DPR approach, more than half of the models improved their results compared to the BM25 model. However, when using the relevance classification approach, only one model improved the results.

To address **RQ5.3**, which is related to the performance of the hybrid approach compared to BM25 in CA, specifically in the Holy Quran, I followed these steps. First, I needed to determine the number of passages retrieved by BM25. I conducted three experiments for the hybrid approach, which involved using BM25 to retrieve the top passages and then re-ranking them using the AraBERT Base model as a DPR approach. In these three experiments, I fixed all parameters except the number of passages selected from the BM25 list. In the first experiment, the number was set to 100; in the second, it was 50; and in the third, it was 20.

| | Bi-Encoder | | Cross-Encoder | |
|---|---|---|---|---|
| **Model** | **MAP@10** | **MRR@10** | **MAP@10** | **MRR@10** |
| AraBERT Base | 0.244 | 0.413 | 0.077 | 0.131 |
| ArabicBERT | 0.172 | 0.281 | 0.119 | 0.199 |
| QARiB | 0.164 | 0.305 | 0.096 | 0.183 |
| ARBERT | 0.161 | 0.300 | 0.114 | 0.195 |
| MARBERT | 0.154 | 0.277 | 0.057 | 0.122 |
| CAMeL-BERT | 0.182 | 0.315 | 0.125 | 0.219 |
| AraELECTRA | 0.172 | 0.307 | 0.090 | 0.184 |
| CL-AraBERT | 0.179 | 0.302 | 0.093 | 0.168 |
| **Ensemble All** | 0.219 | 0.360 | 0.098 | 0.173 |
| **Ensemble Best** | 0.171 | 0.288 | 0.139 | 0.220 |

Table 5.1: Results of the DPR (bi-encoder) and relevance classification (cross-encoder) approaches using the Arabic transformer-based model (Training datasets: TREC AyaTEC and QuranTrec, Testing dataset: TREC AyaTEC).

Based on the results shown in Table 5.3, 50 was the value that achieved the best result with 0.227 MAP@10 and 0.365 MRR@10. Therefore, I used this value in the subsequent experiments.

Second, I conducted experiments for the following hybrid approach: ranking using BM25 and then re-ranking using the DPR approach (bi-encoder model). The results in Table 5.4 show that the Arabic pre-trained models, when used as a bi-encoder model in the hybrid approach, achieved better results than BM25 on MAP@10, except for MARBERT and CL-AraBERT. The best model was AraBERT Base, which achieved a 0.227 MAP@10 and 0.365 MRR@10.

Third, I conducted experiments for the hybrid approach that ranked the passages using BM25 and then re-ranked them using the relevance classification (cross-encoder model). From the results shown in Table 5.4 and Table 5.5, I noted that BM25 outperformed the hybrid approach, which used relevance classification (cross-encoder model) as a re-ranker across all the Arabic pre-trained models and the multi-language model.

Thus, the answer to the question **RQ5.3** is that when using the DPR approach in the hybrid approach, six out of eight models improved their results compared

| Model | MAP@10 | MRR@10 |
|---|---|---|
| mmarco-mMiniLMv2-L12-H384-v1 | 0.169 | 0.298 |
| stsb-roberta-base | 0.084 | 0.159 |
| ms-marco-MiniLM-L-12-v2 | 0.107 | 0.192 |
| stsb-distilroberta-base | 0.080 | 0.145 |
| qnli-distilroberta-base | 0.198 | 0.198 |
| ms-marco-TinyBERT-L-2-v2 | 0.088 | 0.174 |
| **Ensemble All** | 0.130 | 0.261 |
| **Ensemble Best** | 0.170 | 0.298 |

Table 5.2: Results of the relevance classification approach using the multilingual transformer-based model (Training datasets: TREC AyaTEC and QuranTrec, Testing dataset: TREC AyaTEC).

| Model | Number of the document | MAP@10 | MRR@10 |
|---|---|---|---|
| AraBERT Base | 100 | 0.168 | 0.276 |
| | 50 | 0.227 | 0.365 |
| | 20 | 0.118 | 0.223 |

Table 5.3: Comparison of the different numbers of passages retrieved from the BM25 in a hybrid approach (Training datasets: TREC AyaTEC and QuranTrec, Testing dataset: TREC AyaTEC).

to the BM25 model. However, BM25 outperformed all models of the relevance classification approach in the hybrid approach.

The final question in this study, **RQ5.4**, was related to whether the transfer learning technique using my new and larger Quran dataset (QuranTrec) and/or MSA datasets and/or the ensemble technique enhanced the performance of using the pre-trained models in the PR task. The question was divided into two parts. The first part concerned training the pre-trained PR model using different datasets, and the results are shown in Table 5.6. The results indicate that training the AraBERT Base model on the two CA datasets (TREC AyaTEC and QuranTrec) while using it as a bi-encoder in the hybrid approach performed best, achieving a 0.168 MAP@10 and 0.276 MRR@10. The second highest result was achieved by the model when trained using the two CA datasets (TREC AyaTEC

| | Bi-Encoder | | Cross-Encoder | |
|---|---|---|---|---|
| **Model** | **MAP@10** | **MRR@10** | **MAP@10** | **MRR@10** |
| AraBERT Base | 0.227 | 0.365 | 0.167 | 0.273 |
| ArabicBERT | 0.198 | 0.328 | 0.155 | 0.290 |
| QARiB | 0.204 | 0.347 | 0.140 | 0.234 |
| ARBERT | 0.222 | 0.405 | 0.105 | 0.227 |
| MARBERT | 0.165 | 0.294 | 0.098 | 0.157 |
| CAMeL-BERT | 0.180 | 0.294 | 0.093 | 0.205 |
| AraELECTRA | 0.219 | 0.373 | 0.093 | 0.164 |
| CL-AraBERT | 0.139 | 0.315 | 0.119 | 0.233 |
| **Ensemble All** | 0.240 | 0.41 | 0.159 | 0.273 |
| **Ensemble Best** | - | - | 0.083 | 0.180 |

Table 5.4: Results of the hybrid approach that retrieved the passage using BM25 and then re-ranked them using the DPR and relevance classification approaches (Training datasets: TREC AyaTEC and QuranTrec, Testing dataset: TREC AyaTEC).

| **Model** | **MAP@10** | **MRR@10** |
|---|---|---|
| qnli-distilroberta-base | 0.116 | 0.207 |

Table 5.5: Results of the hybrid approach that retrieved passages using BM25 then re-ranks with the relevance classification approach (a multi-language transformer-based model) (Training datasets: TREC AyaTEC and QuranTrec, Testing dataset: TREC AyaTEC).

and QuranTrec) and two MSA datasets (ARCD and Arabic SQuAD), resulting in a 0.152 MAP@10 and 0.290 MRR@10. Training the model on only the two MSA datasets achieved the lowest result, with a 0.039 MAP@10 and 0.149 MRR@10.

The second part was related to applying the ensemble technique to the results of the individual models. The results of applying the ensemble technique to the PR models that utilised the DPR approach are shown in Table 5.1, while the results of the pre-trained models that implemented the relevance classification are shown in Table 5.1 and Table 5.2. Additionally, the ensembling results of the PR pre-trained models that used the hybrid approach are shown in Table 5.4

(with the highest values colored in red). The results of applying the ensemble technique to all the models are shown in the row **Ensemble All** of these tables. From Table 5.1, Table 5.2, and Table 5.4, I noted that the use of the ensemble technique did not improve the results, as the outcomes of some individual models were better. For example, the ensemble technique in Table 5.1 achieved a score of 0.219 MAP@10 and 0.360 MRR@10, while the AraBERT Base model alone achieved a score of 0.244 MAP@10 and 0.413 MRR@10.

The only instance in which the ensemble technique improved performance was when it was implemented on the results of bi-encoder models used in the hybrid approach, where it achieved a score of 0.240 MAP@10 and 0.410 MRR@10. Based on these results, I concluded that the very poor performance of the models affected the results of the ensemble technique. Therefore, I decided to study the application of the ensemble technique on the results of only the two best models, as shown in the **Ensemble Best** row in Table 5.1, Table 5.2, and Table 5.4. The **Ensemble Best** did not improve performance, except when the Arabic pre-trained model was used as a cross-encoder, achieving scores of 0.139 MAP@10 and 0.220 MRR@10, as shown in Table 5.1.

| **Model** | **Dataset** | **MAP@10** | **MRR@10** |
|---|---|---|---|
| AraBERT Base | TREC AyaTEC | 0.056 | 0.170 |
| | TREC AyaTEC + QuranTrec | 0.168 | 0.276 |
| | ARCD + Arabic SQuAD | 0.039 | 0.149 |
| | TREC AyaTEC + QuranTrec + ARCD + Arabic SQuAD | 0.152 | 0.29 |

Table 5.6: Comparison of the different datasets used to train the hybrid approach that retrieved the top-100 passages using BM25 then re-ranked using the DPR approach.

The approach that yielded the highest results (0.244 MAP) for the Quran was the DPR method utilizing the AraBERT Base model, which was further fine-tuned on two Classical Arabic datasets: TREC AyaTEC and QuranTrec.

Thus, the answer to question RQ5.4 is that using a larger Quran dataset for training improved the performance of the pre-trained models in PR. Additionally, the ensemble technique sometimes enhanced the results compared to their performance without these techniques, as shown in Table 5.1, Table 5.2, Table 5.4, and Table 5.6.

### 5.6.3 Analysis and Discussion

When I analysed the passages retrieved by one of the models used in the experiments, I noticed several points that are further detailed in Table 5.7. For example, one of the models used in the experiments of this chapter retrieved a passage because it contained the same word as the question, الملائكة (in English, "angels"), although the passage did not answer the question. In other words, it performed a simple matching process. One of the models may have had difficulty retrieving the first golden passage because it did not contain the word الملائكة (in English, "angels"). However, the passage included the phrase بروح القدس ("divine grace"), which is the name of one of the angels, so it is considered the correct answer. Named entity recognition may help with this difficulty. The reason for the model's failure to retrieve the second golden passage may be because the word ملك (in English, "angel") appeared in the passage, not the word الملائكة (in English, "angels"). The word ملك is the singular form of the word الملائكة.

## 5.7 Experiment 2: Answering Questions from the Hadith

### 5.7.1 Datasets

This research utilized four Arabic datasets for the experiments. The first dataset served both training and evaluation purposes, while the other three datasets were exclusively used for training. The questions in the first two datasets were composed in MSA, whereas the corresponding passages were in CA. The last two datasets were entirely written in MSA. Details of each dataset are as follows:

1. **HadithTrec:** This dataset consists of three files: the Hadith Passage Collection (HPC), HadithTrec questions, and query relevance judgments

| | |
|---|---|
| **Question** | **241**- Who are the angels mentioned in the Quran?<br>٢٤١ - من هم **الملائكة** المذكورون في القرآن؟ |
| **Retrieved Passage** | **17:40-44** - Has God chosen to give you sons, and taken for Himself daughters from among the angels? You utter grievous things indeed!...<br>١٧:٤٠-٤٤ - أفأصفاكم ربكم بالبنين واتخذ من **الملائكة** إناثا إنكم لتقولون قولا عظيما... |
| **Gold Passage 1** | **2:87-88** - Remember We gave Moses the Book and sent after him many an apostle; and to Jesus, son of Mary, We gave clear evidence of the truth, reinforcing him with divine grace...<br>٢:٨٧-٨٨ - ولقد آتينا موسى الكتاب وقفينا من بعده بالرسل وآتينا عيسى ابن مريم البينات وأيدناه بروح **القدس** ... |
| **Gold Passage 2** | **32:10-14** - ...Say: "The angel of death appointed over you will take away your soul, then you will be sent back to your Lord"...<br>٣٢:١٠-١٤ - ... قل يتوفاكم **ملك** الموت الذي وكل بكم ثم إلى ربكم ترجعون... |

Table 5.7: Examples of incorrect passages retrieved by one of the models used in the experiments.

(QRels). The HPC comprises 1,950 passages that represent the entire Jami at-Tirmidh book and HAQA. The Jami at-Tirmidh book was divided into sections, with each section addressing a specific topic. Therefore, I collected two consecutive hadiths into one passage. Each passage has an ID number. The question file includes 1,186 questions, each with its own ID. Some long queries were removed because the BERT model variant does not work with long queries. The questions were divided into 85.32% for training, 4.63% for development, and 10.03% for the test sets. The QRels file is composed of 1,383 gold records, with each record being a pairing of a question ID and the ID of the passage that answers it. The distribution of the HadithTrec dataset is shown in Table 5.8.

| | % | **# of the Pairs (Questions & Passages** | **# Questions** |
|---|---|---|---|
| **Training** | 84.1% | 1164 | 1012 |
| **Development** | 4.3% | 60 | 55 |
| **Test** | 11.5% | 159 | 119 |
| **All** | 100% | 1,383 | 1186 |

Table 5.8: The statistics for the number of records in the HadithTrec.

2. **QuranTrec**: Its details are explained in Section 5.6.1.
3. **ARCD**: Its details are explained in Section 5.6.1.
4. **Arabic SQuAD**: Its details are explained in Section 5.6.1.

### 5.7.2 Results

To answer **RQ5.2**, which is related to comparing the performance of the DPR and relevance classification approaches with BM25, I conducted three experiments. First, I built a PR model using BM25. Second, I built a PR model using the DPR approach; those results appear in the "Bi-Encoder" column in Table 5.9. Third, I used the relevance classification approach to build a PR model, and the results from using Arabic models appear in the "Cross-Encoder" column in Table 5.9, while the results from using multilingual models appear in Table 5.10. BM25 obtained 0.307 MAP@10 and 0.315 MRR@10.

From the results, I noted that two out of eight models achieved higher results than BM25 in terms of MAP@10 when used in the DPR approach (results are colored in blue). The models were CAMeL-BERT and CL-AraBERT. The highest-performing model was CAMeL-BERT in terms of MAP@10 and

MRR@10, achieving a 0.400 MAP@10 and 0.425 MRR@10. From Table 5.9, I noted that five out of eight models achieved higher results than BM25 in terms of MAP@10 when used in the relevance classification approach. The models were ArabicBERT, ARBERT, CAMeL-BERT, AraELECTRA, and CL-AraBERT. The best among them was CAMeL-BERT, with a 0.422 MAP@10 and 0.442 MRR@10. There was no model that outperformed BM25 in the multi-language models, as shown in Table 5.10.

| | **Bi-Encoder** | | **Cross-Encoder** | |
|---|---|---|---|---|
| **Model** | **MAP@10** | **MRR@10** | **MAP@10** | **MRR@10** |
| AraBERT Base | 0.290 | 0.309 | 0.298 | 0.326 |
| ArabicBERT | 0.243 | 0.264 | 0.308 | 0.325 |
| QARiB | 0.301 | 0.318 | 0.288 | 0.293 |
| ARBERT | 0.247 | 0.268 | 0.377 | 0.399 |
| MARBERT | 0.247 | 0.268 | 0.280 | 0.291 |
| CAMeL-BERT | 0.400 | 0.425 | 0.422 | 0.442 |
| AraELECTRA | 0.188 | 0.200 | 0.343 | 0.353 |
| CL-AraBERT | 0.400 | 0.422 | 0.412 | 0.428 |
| **Ensemble All** | 0.395 | 0.418 | 0.376 | 0.393 |
| **Ensemble Best** | 0.406 | 0.427 | 0.411 | 0.426 |

Table 5.9: Results of the DPR and relevance classification approaches using the Arabic transformer-based model (Training datasets:HadithTrec, QuranTrec, SQuAD, and ARCD, Testing dataset: HadithTrec).

From the above, the answer to question **RQ5.2** is that when using the DPR approach, two models improved their results compared to the BM25 model, while when using the relevance classification approach, more than half of the models performed better.

To address **RQ5.3**, which is related to the performance of the hybrid approach compared to BM25 in CA, specifically in Hadith, I followed these steps. First, I needed to determine the number of passages retrieved by BM25. I conducted three experiments for the hybrid approach, which involved using BM25 to retrieve the top passages and then re-ranking them using the AraBERT Base model as a DPR approach. In these three experiments, I fixed all parameters except the

| Model | MAP@10 | MRR@10 |
|---|---|---|
| stsb-roberta-base | 0.078 | 0.080 |
| ms-marco-MiniLM-L-12-v2 | 0.275 | 0.284 |
| stsb-distilroberta-base | 0.199 | 0.210 |
| qnli-distilroberta-base | 0.305 | 0.318 |
| ms-marco-TinyBERT-L-2-v2 | 0.286 | 0.298 |
| **Ensemble All** | 0.300 | 0.312 |
| **Ensemble Best** | 0.292 | 0.304 |

Table 5.10: Results of the relevance classification approach using the multilingual transformer-based model (Training datasets: HadithTrec, QuranTrec, SQuAD, and ARCD, Testing dataset: HadithTrec).

number of passages selected from the BM25 list. In the first experiment, the number was set to 100; in the second, it was 50; and in the third, it was 20. Based on the results shown in Table 5.11, 100 was the value that achieved the best result with 0.384 MAP@10 and 0.402 MRR@10. Therefore, I used this number in the upcoming experiments.

Second, I conducted experiments for the following hybrid approach: ranking using BM25 and then re-ranking using the DPR approach (bi-encoder model). The results in Table 5.12 show that the Arabic pre-trained models, when used as a bi-encoder model in the hybrid approach, achieved better results than BM25 on MAP@10, except for ArabicBERT and AraELECTRA. The best model was CAMeL-BERT, which achieved a 0.410 MAP@10 and 0.428 MRR@10.

Third, I conducted experiments for the hybrid approach that ranked the passages using BM25 and then re-ranked them using the relevance classification (cross-encoder model). From the results shown in Table 5.12 and Table 5.13, I noted that six out of eight Arabic pre-trained models achieved higher results than BM25 in terms of MAP@10 when used in the relevance classification approach. The models were AraBERT Base, ArabicBERT, ARBERT, CAMeL-BERT, AraELECTRA, and CL-AraBERT. BM25 outperformed all the multi-language models.

Thus, the answer to the question **RQ5.3** is that when using the DPR or relevance classification approach in the hybrid method, six out of eight Arabic

| Model | Number of the document | MAP@10 | MRR@10 |
|---|---|---|---|
| AraBERT Base | 100 | 0.384 | 0.402 |
| | 50 | 0.284 | 0.296 |
| | 20 | 0.284 | 0.296 |

Table 5.11: Comparison of the different numbers of passages retrieved from the BM25 in a hybrid approach (Training datasets:HadithTrec, QuranTrec, SQuAD, and ARCD, Testing dataset: HadithTrec).

| | Bi-Encoder | | Cross-Encoder | |
|---|---|---|---|---|
| **Model** | **MAP@10** | **MRR@10** | **MAP@10** | **MRR@10** |
| AraBERT Base | 0.384 | 0.402 | 0.356 | 0.375 |
| ArabicBERT | 0.243 | 0.264 | 0.308 | 0.325 |
| QARiB | 0.345 | 0.361 | 0.288 | 0.293 |
| ARBERT | 0.309 | 0.328 | 0.377 | 0.399 |
| MARBERT | 0.309 | 0.328 | 0.280 | 0.291 |
| CAMeL-BERT | 0.410 | 0.428 | 0.422 | 0.442 |
| AraELECTRA | 0.215 | 0.228 | 0.343 | 0.353 |
| CL-AraBERT | 0.400 | 0.422 | 0.392 | 0.405 |
| **Ensemble All** | 0.399 | 0.412 | 0.376 | 0.393 |
| **Ensemble Best** | 0.408 | 0.429 | 0.411 | 0.426 |

Table 5.12: Results of the hybrid approach that retrieved the passage using BM25 and then re-ranked them using the DPR and relevance classification approaches (Training datasets:HadithTrec, QuranTrec, SQuAD, and ARCD, Testing dataset: HadithTrec).

pre-trained models improved their results compared to the BM25 model.

The final question in this study, **RQ5.5**, was related to whether the transfer learning technique and/or the ensemble technique enhanced the performance of the pre-trained models in the PR of Hadith. The question was divided into two parts. The first part concerns training the pre-trained PR model using different datasets, and the results are shown in Table 5.14. The results show that training the AraBERT Base model on the two CA datasets (TREC AyaTEC and QuranTrec) and the two MSA datasets (ARCD and Arabic SQuAD) while used as

| Model | MAP@10 | MRR@10 |
|---|---|---|
| stsb-roberta-base | 0.078 | 0.080 |
| ms-marco-MiniLM-L-12-v2 | 0.275 | 0.284 |
| stsb-distilroberta-base | 0.199 | 0.210 |
| qnli-distilroberta-base | 0.305 | 0.318 |
| ms-marco-TinyBERT-L-2-v2 | 0.286 | 0.298 |

Table 5.13: Results of the hybrid approach that retrieved passages using BM25 then re-ranks with the relevance classification approach (a multi-language transformer-based model) (Training datasets:HadithTrec, QuranTrec, SQuAD, and ARCD, Testing dataset: HadithTrec).

a bi-encoder in the hybrid approach performed best, achieving a 0.384 MAP@10 and 0.402 MRR@10. The second highest result was achieved by the model when trained using the two CA datasets (TREC AyaTEC and QuranTrec), with a 0.384 MAP@10 and 0.395 MRR@10. Training the model on only the two MSA datasets yielded the lowest result, with a 0.126 MAP@10 and 0.134 MRR@10.

The second part is related to applying the ensemble technique to the results of the individual models. The results of applying the ensemble technique to the PR models that implemented the DPR approach are shown in Table 5.9, while the results of the pre-trained models that implement the relevance classification are shown in Table 5.9 and Table 5.10. Additionally, the ensembling results of the PR pre-trained models that used the hybrid approach are shown in Table 5.12 (with the highest values colored in red). The results of applying the ensemble technique to all the models are displayed in the row **Ensemble All** of these tables.

From Table 5.9, Table 5.10, and Table 5.12, I noted that using the ensemble technique did not improve the results since the results of some individual models were better. For example, the ensemble technique in Table 5.9 achieved a score of 0.395 MAP@10 and 0.418 MRR@10, while the CAMeL-BERT model alone received a score of 0.400 MAP@10 and 0.425 MRR@10. Based on these results, I conclude that the very poor performance of the models affects the results of the ensemble technique. Therefore, I decided to study the application of the ensemble technique on the results of only the two best models, as shown in the **Ensemble Best** row in Table 5.9, Table 5.10, and Table 5.12. The **Ensemble**

**Best** improved performance when the Arabic pre-trained model was used as a bi-encoder, achieving scores of 0.406 MAP@10 and 0.427 MRR@10, as shown in Table 5.9.

| Model | Dataset | MAP@10 | MRR@10 |
|---|---|---|---|
| AraBERT Base | HadithTrec | 0.336 | 0.358 |
| | HadithTrec + QuranTrec | 0.384 | 0.395 |
| | ARCD + Arabic SQuAD | 0.126 | 0.134 |
| | HadithTrec + QuranTrec + ARCD + Arabic SQuAD | 0.384 | 0.402 |

Table 5.14: Comparison of the different datasets used to train the hybrid approach that retrieved the top-100 passages using BM25 then re-ranked using the DPR approach.

Two models demonstrated the best performance on the Hadith dataset: the hybrid method that employed the CAMeL-BERT model and the relevance classification method, which also utilized the CAMeL-BERT model. Both methods achieved a MAP@10 score of 0.422. The CAMeL-BERT model used in both approaches was trained on two Classical Arabic datasets (HadithTrec and QuranTrec), as well as two Modern Standard Arabic datasets (ARCD and Arabic SQuAD).

Thus, the answer to the question **RQ5.5** is that using transfer learning enhanced the performance of the pre-trained models in the PR of Hadith, while the ensemble technique sometimes improved the results.

### 5.7.3 Analysis and Discussion

When examining the model's answers, two points were noted that are further detailed in Table 5.15 and Table 5.16. First, one of the models performed a simple matching process. In the first example, it retrieved a passage because it included the same question terms صلاة، صلى الله عليه وسلم، الظهر (in English, "Salat, 'may God bless him and grant him peace,' Az-Zuhr"), even though the passage did not answer the question. In the second example, it retrieved a passage because it included the same question terms صلاة (in English, "Salat").

Second, the model's failure to retrieve the golden passage may be because the word in the passage is synonymous with the word in the question. For example, the word يستغفر (in English, "forgiving") appeared in the passage, not the word التوبة (in English, "repentance") that appeared in the question.

## 5.8 Answering the Research Questions

Based on the experiments conducted in this chapter, the following questions can be answered:

- **RQ2:** Can PLMs answer questions from the Quran and Hadith in Arabic? The results showed an acceptable performance of the models but indicated a need for improvement. Our findings may open the door for more research on PR tasks in Arabic and other languages.

- **RQ3:** Does the size of the new Quran question-answer dataset affect the performance of the pre-trained transformer models? Results showed that it improved the performance of the PR task.

## 5.9 Conclusion

One of the primary objectives of this thesis was to develop a model capable of answering questions from the Holy Quran and Hadith using transformer-based models. The model was designed to handle two key tasks, one of which was passage retrieval (PR).

This chapter began with a literature review of existing PR approaches. I then evaluated the performance of transformer-based models in dense passage retrieval (DPR), relevance classification, and hybrid approaches. Finally, I compared the performance of these approaches with BM25.

When applying the DPR approach to the Quran PR model, more than half of the models outperformed BM25, demonstrating the potential of dense retrieval in

| | |
|---|---|
| **Question** | **1113**- Why did the Prophet, may God bless him and grant him peace, regularly perform the voluntary Salat Az-Zuhr? **١١١٣** - ما سبب مواظبة النبي صلى الله عليه و سلم على راتبة صلاة الظهر ؟ |
| **Retrieved Passage** | **2:274-274** - ... I asked Aishah about the Salat of Allah's Messenger (may God bless him and grant him peace). She said: 'He would pray two Rak'ah before Az-Zuhr and two Rak'ah after it, and two after Al-Maghrib, and two Rak'ah after Al-Isha, and two before Al-Fajr.... **٢:٢٧٤-٢٧٤** - ... سألت عائشة عن صلاة رسول الله صلى الله عليه وسلم فقالت كان يصلي قبل الظهر ركعتين وبعدها ركعتين وبعد المغرب ثنتين وبعد العشاء ركعتين وقبل الفجر ثنتين ... |
| **Gold Passage** | **1:421-421** - ... Umm Habiba said she heard God's Messenger say, "If anyone keeps on praying four rak'as before and four after the noon prayer, God will forbid that he be sent to hell." ... **١:٤٢١-٤٢١** - ... سمعت رسول الله صلى الله عليه وسلم يقول من حافظ على أربع ركعات قبل الظهر وأربع بعدها حرمه الله على النار ... |

Table 5.15: Examples of incorrect passages retrieved by one of the models part 1.

Quranic PR. For the Hadith PR model, the DPR approach improved retrieval results in two models compared to BM25. The model that achieved the highest score was CAMeL-BERT, with a MAP@10 of 0.400 and an MRR@10 of 0.425 when

| | |
|---|---|
| **Question** | **1174**- What is the evidence for the prayer of repentance?<br>١١٧٤ - ما دليل **صلاة** التوبة؟ |
| **Retrieved Passage** | **1:146-146** - I was with Buraida on a cloudy day and he said, "Offer the Asr prayer earlier as the Prophet (may God bless him and grant him peace) said, Whoever leaves the Asr prayer will have all his (good) deeds annulled." ...<br>١:١٤٦-١٤٦ - حديث بريدة رضي الله عنه في يوم ذي غيم فقال بكروا **بالصلاة** فإن النبي صلى الله عليه وسلم قال من ترك **صلاة** العصر حبط عمله ... |
| **Gold Passage** | **1:443-443** - ... he heard God's Messenger say, "No man will commit a sin, then get up and purify himself, then pray, then ask God's forgiveness without God forgiving him." ...<br>١:٤٤٣-٤٤٣ - ... سمعت رسول الله صلى الله عليه وسلم يقول ما من رجل يذنب ذنبا ثم يقوم فيتطهر ثم يصلي ثم يستغفر الله إلا غفر الله له ... |

Table 5.16: Examples of incorrect passages retrieved by one of the models part 2.

used for Hadith PR, while BM25 obtained 0.307 MAP@10 and 0.315 MRR@10.

Furthermore, the relevance classification approach was evaluated against BM25 to assess its effectiveness. However, when applying this approach to the Quran PR model, only one model showed an improvement over BM25, suggesting that relevance classification may not be as effective for Quranic PR. In contrast, this approach proved to be more effective for Hadith PR, improving performance in more than half of the models. The best-performing model was CAMeL-BERT, achieving a MAP@10 of 0.422 and an MRR@10 of 0.442 when answering Hadith-related questions.

The best result in the hybrid approach was a 0.422 MAP when using CAMeL-BERT as a cross-encoder.

I tried to improve the performance of these models by further fine-tuning the models using an additional dataset (including our dataset) or using the ensemble technique. Additionally, transfer learning proved to be a valuable optimization technique, while the ensemble method showed mixed results.

# Chapter 6

# Evaluating the Pre-trained Language Models (PLMs) for Machine Reading Comprehension Task

## 6.1 Introduction

Studies have shown the promising performance of transformer-based models in many NLP tasks, including MRC. However, the performance of these models has not been explored extensively or deeply in CA texts. CA is considered a low-resource language due to the availability of only one small dataset (the QRCD) specific to the Quran (Kazi *et al.*, 2023). To fill this gap, I first created two datasets: one for the Quranic text (QMRC) and the other for the Hadith (HMRC), which are considered the main sources of Islam. Their text requires a deep understanding of the content, and their terminology can have an interpretation that is different from that of other books. Second, I explored the performance of all the pre-trained transformer-based models available for the Arabic language when used as answering models for Quran and Hadith questions. Third, I studied the impact of further training the models on a CA dataset, such as the Quran dataset and/or MSA datasets. Finally, I selected the best-performing models and applied the ensemble method to their results.

This chapter is built on two publications:

- Alnefaie S, Atwell E, Alsalka MA. Question Answering Over the Arabic Hadith Using Transformer Models **Accepted in** The Eighth International Conference on Arabic Language Processing, (ICALP 2024).
- (Alnefaie *et al.*, 2023a) received the Best Paper award.

## 6.2 Background

MRC is one of the key NLP tasks. In this task, the MRC model takes a text passage and a question as inputs and then extracts the correct answer from the passage (Baradaran *et al.*, 2022). The MRC model architecture is shown in Figure 6.1.

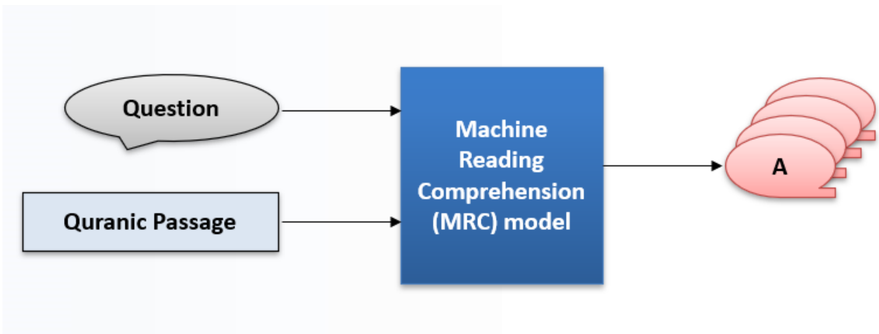



Figure 6.1: Machine reading comprehension model architecture.

## 6.3 Related Work

Recently, some studies have focused on the performance of pre-trained models for MRC with CA text. They used the QRCD for training and testing. This dataset consists of questions and answers derived from the text of the Holy Quran. The main differences between these studies were the models used and the methods of improvement. Different Arabic BERT models have been employed in this research, such as AraBERT, ARBERT, MARBERT, and QARiB. The optimization methods they used can be divided into two types: employing transfer learning and applying an ensemble approach to the different models.

- **Transfer Learning Approach**

  MSA datasets were used in some studies to further fine-tune the model, in addition to the QRCD. Mostafa & Mohamed (2022) proposed using three MSA datasets to train the AraELECTRA model: the Arabic Reading Comprehension Dataset (ARCD), Ar-TyDi, and Arabic-SQuAD. Their system achieved scores of 0.23, 0.54, and 0.52 in exact match (EM), partial reciprocal ranking (pRR), and F1@1, respectively. Additionally, Malhas & Elsayed (2022) conducted three experiments. First, only the Arabic-SQuAD and ARCD datasets, which are MSA datasets, were used to fine-tune the AraBERT and CL-AraBERT models. Second, only QRCD was used to fine-tune the two models. Third, the ARCD and Arabic-SQuAD datasets were used to fine-tune the two models, and then the models were further fine-tuned on the QRCD. The highest results were achieved by AraBERT and CL-AraBERT when they were fine-tuned using both the MSA and QRCD datasets: with partial average precision (paP)@10 scores of 49.53 and 53.28, respectively.

  Other studies have used a CA dataset to fine-tune the model. There is no MRC dataset currently available for CA other than the QRCD, so they used the questions and answers dataset and formatted it appropriately for MRC. Wasfey *et al.* (2022) suggested using the AQQAC dataset to fine-tune the AraBERTv02Base model. The results showed a 0.25 EM, 0.5 F1@1, and 0.52 pRR. Aftab & Malik (2022) proposed fine-tuning the baseline BERT using the AQQAC and QRCD. Their model achieved a 0.30 pRR, 0.26 F1@1, and 0.08 EM.

- **Ensemble Approach**

  ElKomy & Sarhan (2022) developed an MRC system for the Arabic Quran by training five BERT models using the QRCD. The models were AraBERT Large, AraBERT Base, ARBERT, MARBERT, and QARiB. They obtained answers for the test set of the QRCD using each model individually. Then, they applied an ensemble approach to these models to achieve better results. This system achieved a 0.50 F1@1, 56.6 pRR, and 26.8 EM.

To the best of our knowledge, no study has been conducted on the impact of combining the following three factors in building the Arabic Quran MRC model: First, Arabic pre-trained models are fine-tuned using CA and MSA datasets. Second, an ensemble approach is applied to the results using the majority vote. Finally, the final list is refined through several post-processing steps. The research questions in this chapter are:

1. **RQ6.1** Does combining fine-tuning the models with a large CA dataset and MSA dataset improve the results?
2. **RQ6.2** Does ensembling further fine-tuned models and subsequently applying post-processing steps improve the results?

To the best of our knowledge, there is no existing study that has built an MRC model for the Hadith. In addition, no study could be found that has explored the CA language in depth. The research questions in this study were as follows:

1. **RQ6.3**: Do the pre-trained models perform well when used in the MRC task to answer questions about the Hadith?
2. **RQ6.4**: Does further fine-tuning of the models with large MSA and/or CA datasets enhance their performance?
3. **RQ6.5**: Does applying the ensemble approach and then implementing post-processing steps for the pre-trained models improve the results for the Hadith MRC?

These questions are numbered starting from 6 because they are in Chapter 6 and are not related to the general questions.

## 6.4 The Proposed Models

The pre-trained transformer-based models formed the basis of my methodology. I conducted two separate experiments: the first experiment focused on finding answers from the Holy Quran, while the second focused on finding answers from the Hadith. As the first step in Experiment 1, I fine-tuned all available Arabic pre-trained models with the QRCD_v1.2 training set. The first step in Experiment 2 involved using the HMRC training set to fine-tune the nine Arabic transformer-based models. The nine Arabic pre-trained models were: ArabicBERT (Safaya *et al.*, 2020), AraBERT Large, AraBERT Base (Antoun *et al.*, 2020), ARBERT, MARBERT (Abdul-Mageed *et al.*, 2021), QARiB (Abdelali *et al.*, 2021), Ara-ELECTRA (Antoun *et al.*, 2021), CAMeL-BERT (Inoue *et al.*, 2021), and CL-AraBERT (Malhas & Elsayed, 2022). In Experiment 1, the test dataset was QRCD_v1.2, while in the second experiment, the dataset was HMRC. For my experiments, I set the batch size to 8 for AraBERT Large and 16 for the rest of the models, the number of epochs to 4, and the learning rate to 1e-4. I attempted to improve performance using the following three optimization approaches:

- **Transfer Learning:** I conducted three experiments using this approach. I further fine-tuned the models using different datasets.
  - **Quran Experiment:** I used the CA dataset QMRC. Second, the MSA dataset ARCD was used. Finally, a combination of the QMRC dataset and ARCD was used only for the models that showed an improvement in performance when using one of these two datasets individually.
  - **Hadith Experiment:** The ARCD, considered an MSA dataset, was used to fine-tune all models. Second, the QMRC, considered a CA dataset, was used to fine-tune all models. Third, I identified the models that showed improvement in both previous experiments. I then fine-tuned these models using both the ARCD and QMRC datasets.
- **Ensemble Approach:** I used majority voting among the models to produce the final ranked-list results. I took the top 20 answers with their

scores for each question from each model. I then computed the total score for each answer, which was the sum of the scores obtained from all models. After that, I sorted the answers for each question based on the total score. Finally, I adopted the top 10 answers as the final ranked list.

- **Post-Processing:** There were two issues when producing the ranked list: uninformative answers (as shown in Example 1) and overlapping answers (as shown in Example 2). An uninformative answer is one that contains only stop words. The first issue was resolved by removing these answers from the list. The second was addressed by applying a redundancy elimination algorithm (ElKomy & Sarhan, 2022).

  The input to the redundancy elimination algorithm consists of the passage and the list of initially retrieved answers from the question-answering system. The output is the list of final retrieved answers.

  The idea of this algorithm is to build an array of zeros whose length is equal to the length of the passage. This array is called the "seen" array. Each word in the passage corresponds to a zero. A zero indicates that the word has not yet appeared in the final list of answers. At the beginning, the final list of answers is initialized as an empty list.

  The algorithm processes each answer in the initial list of retrieved answers as follows:

  1. It takes an answer and determines the start and end positions of the answer in the passage.
  2. Using the start and end indexes from Step 1, it extracts a sub-array from the "seen" array that corresponds to the answer.
  3. If the sub-array contains any zeros, it means that the words corresponding to these zeros have not yet been included in the final list of answers.
     - Extract all zero sequences in the sub-array and the sequences of words that correspond to them.
     - For each sequence:

  * Determine the locations of the zero sequence in the "seen" array.
  * Mark the indexes of this sequence with the number 1 in the "seen" array.
  * Add the sequence of words to the list of final retrieved answers.

- **Example 1:**
  * **"pq_id":** 13:18-24_360
  * **"passage":**

للذين استجابوا لربهم الحسنى والذين لم يستجيبوا له لو أن لهم ما في الأرض جميعا ومثله معه لافتدوا به أولئك لهم سوء الحساب ومأواهم جهنم وبئس المهاد. أفمن يعلم أنما أنزل إليك من ربك الحق كمن هو أعمى إنما يتذكر أولو الألباب . الذين يوفون بعهد الله ولا ينقضون الميثاق . والذين يصلون ما أمر الله به أن يوصل ويخشون ربهم ويخافون سوء الحساب . والذين صبروا ابتغاء وجه ربهم وأقاموا الصلاة وأنفقوا مما رزقناهم سرا وعلانية ويدرؤون بالحسنة السيئة أولئك لهم عقبى الدار . جنات عدن يدخلونها ومن صلح من آبائهم وأزواجهم وذرياتهم والملائكة يدخلون عليهم من كل باب . سلام عليكم بما صبرتم فنعم عقبى الدار .

"For those who respond to their Lord, are (all) good things. But those who respond not to Him,- Even if they had all that is in the heavens and on earth, and as much more, (in vain) would they offer it for ransom. For them will the reckoning be terrible: their abode will be Hell,- what a bed of misery! (18) Is then one who doth know that that which hath been revealed unto thee from thy Lord is the Truth, like one who is blind? It is those who are endued with understanding that receive admonition;- (19) Those who fulfil the covenant of Allah and fail not in their plighted word; (20) Those who join together those things which Allah hath commanded to be joined, hold their Lord in awe, and fear the terrible reckoning; (21) Those who patiently persevere, seeking the countenance of their Lord; Establish regular prayers; spend, out of (the gifts) We have bestowed for their sustenance, secretly and openly; and turn off Evil with good: for such there is the final attainment of the (eternal) home,- (22) Gardens of perpetual bliss: they shall enter there, as well as the righteous among their fathers, their spouses, and their offspring: and angels shall enter unto them from every gate (with the salutation): (23) "Peace unto you for that ye persevered in patience! Now how excellent is the final home!" (24)"

* **"question":**

هل سيجمع الله بين المؤمنين وأبنائهم وأهلهم في الجنة ؟

Will God bring the believers together with their children and families in Paradise?

* **"Predicted answer":**

في

on

* **"rank":** 1
* **"score":** 0.1957549469953647
* **"Gold answer":**

جنات عدن يدخلونها ومن صلح من آبائهم وأزواجهم وذرياتهم والملائكة يدخلون عليهم من كل باب

"Gardens of perpetual bliss: they shall enter there, as well as the righteous among their fathers, their spouses, and their offspring: and angels shall enter unto them from every gate (with the salutation): "

– **Example 2:**

* **"pq_id":** "2:177-177_419"

∗ **"question":**

هل احترم الإسلام الأنبياء ؟

Does Islam respect the prophets?

1. **"Predicted answer":**

   **من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين**

   "to believe in Allah and the Last Day, and the Angels, and the Book, and the Messengers"

   **"rank":** 1, ...

2. **"Predicted answer":**

   **البر من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين**

   "it is righteousness- to believe in Allah and the Last Day, and the Angels, and the Book, and the Messengers"

   **"rank":** 2 ...

3. **"Predicted answer":**

   **آمن بالله واليوم الآخر والملائكة والكتاب والنبيين**

   "believe in Allah and the Last Day, and the Angels, and the Book, and the Messengers"

   **"rank":** 3 ...

**- Experiment 1:** The QRCD was the testing dataset. The QRCD, QMRC, and ARCD were the fine-tuning datasets.

**- Experiment 2:** The HMRC was the testing dataset. The HMRC, QMRC, and ARCD were the fine-tuning datasets.

## 6.5 Evaluation

In this study, I used four evaluation metrics: EM, partial reciprocal rank (pRR), partial average precision (pAP), and F1 score for these experiments because all research in the field of the Quran has utilized them. The main evaluation metric was pAP (Malhas & Elsayed, 2020, 2022).

- **EM:** EM is a binary score (0 or 1). If one of the gold answers fully matches the top candidate's answer. Then the EM equals 1.

- **F1:**
  Let R be systems's retrieved ranked list of answers to the question. R=[$r_1$,$r_2$,$r_3$,$r_4$]
  Let A be the set of the gold direct answers to the question.
  A=[$a_1$,$a_2$,$a_3$,$a_4$,$a_5$]
  To calculate the value of F1 we need to calculate the following values:

  1. **For each question:**

     (a) **Compute the $m_r$ for each system answer:** I define the answer-matching score ( m ) for a system's response ( r ), represented as $m_r$, as the highest matching score of ( r ) compared to all direct gold standard answers ( A ) for the given question.

     $$m_r = max_{a \epsilon A} f_m(r, a) \tag{6.1}$$

     where $f_m(r, a)$ is the F1 measure of the system answer r and the gold answer a.

     $$f_m(r, a) = F1(V_r|V_a) \tag{6.2}$$

     where $V_r$ and $V_a$ represent part of verses (R) and the gold standard answers (A), respectively.
     $R$=[$r_1$,$r_2$,$r_3$,$r_4$]
     $A$=[$a_1$,$a_2$,$a_3$,$a_4$, $a_5$]
     $m_{r_i} = max(F1(r_i|a_1), F1(r_i|a_2), F1(r_i|a_3), F1(r_i|a_4), F1(r_i|a_5))$
     where $i$ from 1 to 4.

     $$F1(r_i|a_j) = \frac{2 * Precision(r_i a_j) * Recall(r_i a_j)}{Precision(r_i a_j) + Recall(r_i a_j)} \tag{6.3}$$

     where $i$ from 1 to 4 and $j$ from 1 to 5.

     $$Precision(r_i a_j) = \frac{\text{retrieved and relevant words}}{\text{all retrieved words}} \tag{6.4}$$

$$Recall(r_i a_j) = \frac{\text{retrieved and relevant words}}{\text{all relevant words}} \tag{6.5}$$

For example, Let R1 be system 1's retrieved ranked list of answers to the question.
R1=[$r_1$= الكافرين "who reject Faith",
$r_2$=المعتدين "transgressors",
$r_3$= الذين يقاتلون في سبيله صفا كأنهم بنيان مرصوص "who fight in His Cause in battle array, as if they were a solid cemented structure."]

Let R2 be system 2's retrieved ranked list of answers to the question.
R2=[ $r_1$= قل إن كنتم تحبون الله فاتبعوني يحببكم الله "If you do love Allah, Follow me: Allah will love you."
$r_2$= الكافرين "who reject Faith",
$r_3$=المعتدين "transgressors"]

Let A be the set of the gold direct answers to the question.
A=[$a_1$= الكافرين "who reject Faith",
$a_2$=المعتدين "transgressors"]

- **System 1's:** I treat verses as if they were a bag of words. Compute $m_{r_1}, m_{r_2}, m_{r_3}$:
  $m_{r_i} = max(F1(r_i|a_1), F1(r_i|a_2))$

  $$F1(r_i|a_j) = \frac{2 * Precision(r_i a_j) * Recall(r_i a_j)}{Precision(r_i a_j) + Recall(r_i a_j)} \tag{6.6}$$

  where $i$ from 1 to 3 and where $j$ from 1 to 2.
  1-Compute $m_{r_1}$:
  $m_{r_1} = max(F1(r_1|a_1), F1(r_1|a_2))$
  1.1-Compute $F1(r_1|a_1)$:
  $r_1$ = الكافرين "who reject Faith"

$a_1$= الكافرين "who reject Faith"
Fully match.

Precision $(r_1 a_1)$ =

$$\frac{\text{retrieved and relevant words one word الكافرين from } r_1}{\text{all retrieved words one word الكافرين from } r_1} = \frac{1}{1} = 1 \quad (6.7)$$

Recall $(r_1 a_1)$ =

$$\frac{\text{retrieved and relevant verses (one word الكافرين from } r_1}{\text{all relevant verses (one word الكافرين from } a_1} = \frac{1}{1} = 1 \quad (6.8)$$

$$F1(r_1|a_1) = \frac{2 * 1 * 1}{1 + 1} = \frac{2}{2} = 1 \quad (6.9)$$

$F1(r_1|a_1) = 1$
1.2-Compute $F1(r_1|a_2)$:
$r_1$ = الكافرين "who reject Faith"
$a_2$= المعتدين "transgressors"
$F1(r_1|a_2) = 0$
No match.
$m_{r_1} = max(1, 0) = 1$
$a_1$ is removed since it was matched with $r_1$.

2-Compute $m_{r_2}$:
$m_{r_2} = max(F1(r_2|a_2))$
2.1-Compute $F1(r_2|a_2)$:

$r_2 =$ المعتدين "transgressors"
$a_2$= المعتدين "transgressors"
$F1(r_2|a_2) = 1$
Fully match.
$m_{r_2} = max(1) = 1$
$a_2$ is removed since it was matched with $r_2$.

3-Compute $m_{r_3}$:
No remaining gold answers to match
$m_{r_3} = 0$

**System 2's:** I treat verses as if they were a bag of words. Computer $m_{r_1}, m_{r_2}, m_{r_3}$:
1-Compute $m_{r_1}$:
$m_{r_1} = max(F1(r_1|a_1), F1(r_1|a_2))$
1.1-Compute $F1(r_1|a_1)$:
$r_1 =$ قل إن كنتم تحبون الله فاتبعوني يحببكم الله "If you do love Allah, Follow me: Allah will love you."
$a_1$= الكافرين "who reject Faith"
$F1(r_1|a_1) = 0$
No match.
1.2-Compute $F1(r_1|a_2)$:
$r_1 =$ قل إن كنتم تحبون الله فاتبعوني يحببكم الله "If you do love Allah, Follow me: Allah will love you."
$a_2$= المعتدين "transgressors"
$F1(r_1|a_2) = 0$
No match.
$m_{r_1} = max(0, 0) = 0$

2-Compute $m_{r_2}$:
$m_{r_2} = max(F1(r_2|a_1), F1(r_2|a_2))$
2.1-Compute $F1(r_2|a_1)$:
$r_2 =$ الكافرين "who reject Faith"

$a_1$= الكافرين "who reject Faith"
$F1(r_2|a_1) = 1$
Fully match.
2.2-Compute $F1(r_2|a_2)$:
$r_2$ = الكافرين "who reject Faith",
$a_2$= المعتدين "transgressors"
$F1(r_2|a_2) = 0$
No match.
$m_{r_2} = max(1, 0) = 1$
$a_1$ is removed since it was matched with $r_2$.

3-Compute $m_{r_3}$:
$m_{r_3} = max(F1(r_3|a_2))$
3.1-Compute $F1(r_3|a_2)$:
$r_3$ = المعتدين "transgressors"
$a_2$= المعتدين "transgressors"
$F1(r_3|a_2) = 1$
Fully match.
$m_{r_3} = max(1) = 1$
$a_2$ is removed since it was matched with $r_3$.

(b) **Compute the overall $F1$ for the question:**

$$F1 = \frac{\sum_{i=1}^{n} m_{r_i}}{n} \tag{6.10}$$

$n$ is the number of the system answers. In the above example 1 $n$=4 and $n$=3 in example 2 .

2. **Compute the $F1$ for all question:**

$$F1 = \frac{\sum_{w=1}^{Q} F1_w}{Q} \tag{6.11}$$

Where $Q$ is the number of questions in the test set.

- **F1@1:**

$$F1@1 = m_{r_1} \tag{6.12}$$

F1@1 is computed only using the first system answer $r_1$.

- **pRR:**

$$pRR = \frac{m_{r_k}}{k}; k = min\{k|m_{r_k} > 0\} \tag{6.13}$$

The $k$ is the location or the rank of the first answer in the system list R that partially matches the golden answers, i.e. it has an $m_r$ score greater than zero.

- **pAP:**

$$pAP = \frac{1}{|A|} \sum_{K=1}^{|R|} 1\{m_{r_k} > 0\} pPrec@K(R) \tag{6.14}$$

where the number of answers in the ranked list is represented by |R|, and the number of gold answers is represented by |A|. $1\{m_{r_k} > 0\}$ is a function is set to 1 if the predicted answer at position (K) fully or partially matches one of the gold answers, and set to 0 if it does not match either fully or partially. Partial Precision at rank (K) is computed as follows:

$$pPrec@K(R) = \frac{1}{K} \sum_{i=1}^{K} m_{r_i} \tag{6.15}$$

where ( R ) represents the ranked list of predicted answers returned by the system, ( $r_i$ ) denotes the predicted answer at a position or rank ($i$) in ( $R$), and ($m_{r_i}$) refers to the partial matching score of ( $r_i$), as in Eq. 6.1.

For the same previous Example,
Let R1 be the system 1's retrieved ranked list of answers to the question.
I computed:
$m_{r_1} = 1$ $m_{r_2} = 1$ $m_{r_3} = 0$

R2 be the system 2's retrieved ranked list of answers to the question.
I computed:
$m_{r_1} = 0$ $m_{r_2} = 1$ $m_{r_3} = 1$
Let A be the set of the gold direct answers to the question.

A=[$a_1$= الكافرين "who reject Faith",
$a_2$=المعتدين "transgressors"]

1. To compute the value of $pRR$ for the first system:

   - **Find the first $m_{r_k}$ not equal 0**
     $m_{r_1} = 1$
   - **Calculate the value of $pRR$ using Eq. 6.13**

$$pRR = \frac{m_{r_1}}{1} = \frac{1}{1} = 1 \tag{6.16}$$

2. To compute the value of $pRR$ for the second system:

   - **Find the first $m_{r_k}$ not equal 0**
     $m_{r_2} = 1$
   - **Calculate the value of $pRR$ using Eq. 6.13**

$$pRR = \frac{m_{r_2}}{2} = \frac{1}{2} = 0.5 \tag{6.17}$$

   The PRR value in the first system is 1 higher than the PRR value in the second system, which is 0.5. This difference occurs because the gold answer in the first system appeared as the first expected answer, while in the second system, it appeared as the second answer.

3. To compute the value of $pAP$ for the first system:

   - **Find all $m_{r_k}$**
     $m_{r_1}$=1
     $m_{r_2}$=1
     $m_{r_3}$=0

- **Calculate the value of $pAP$ using Eq. 6.14 and Eq. 6.15**

$$pAP = \frac{1}{2}\sum_{K=1}^{3} 1\{m_{r_k} > 0\}pPrec@K(R) = \frac{1}{2}(\frac{1}{1} + \frac{1+1}{2} + 0) = \frac{1}{2}(1+1+0) = \frac{1}{2}(2) = 1 \tag{6.18}$$

4. To compute the value of $pAP$ for the second system:

   - **Find all $m_{r_k}$**
     $m_{r_1}$=0
     $m_{r_2}$=1
     $m_{r_3}$=1
   - **Calculate the value of $pAP$ using Eq. 7.16 and Eq. 7.14**

$$pAP = \frac{1}{2}\sum_{K=1}^{3} 1\{m_{r_k} > 0\}pPrec@K(R) = \frac{1}{2}(0 + \frac{0+1}{2} + \frac{0+1+1}{3}) = \frac{1}{2}(\frac{1}{2} + \frac{2}{3}) = \frac{1}{2}(\frac{1}{2} + \frac{2}{3}) = \frac{1}{2}(0.5 + 0.666) = \frac{1}{2}(1.166) = 0.83 \tag{6.19}$$

   pAP of System 1 = 1
   pAP of System 2 = 0.83
   The pAP value in the first system is higher than that in the second system. This is because the gold answers in the first system appeared at higher ranks compared to those in the second system. For more examples of partial matches, please refer to (Malhas & Elsayed, 2022).

Each of these four metrics (EM, F1@1, pRR, pAP@10) was computed for each question. I then calculated the overall score for each metric by averaging it across all questions.

## 6.6 Experiment 1: Answering Questions from the Holy Quran

### 6.6.1 Datasets

In this study, I used three different datasets, as follows:

1. **QRCD:** QRCD_v1.2 consisted of 1,399 question-passage-answer triplets in the training and development splits. It was split into 70%, 10%, and 20% for the training, development, and test sets, respectively (Malhas & Elsayed, 2022).

2. **ARCD**: Its details are explained in Section 5.6.1.

3. **Quran MRC (QMRC):** It is an extended version of the QUQA dataset. The QUQA is a question-answer pairs dataset, while QMRC is a question-passage-answer triplet dataset regarding the Arabic Holy Quran. The QMRC is an MRC dataset. The process of finding the passage for each answer in the QUQA to fit the QMRC was as follows:

   (a) Determined the beginning and end numbers of the verses in the answer.

   (b) The Quran is comprised of a series of verses. I selected the answer verses from the Holy Quran and then extracted the fragment by specifying one verse before the answer verses, the answer verses themselves, and one verse after the answer verses. The passage consists of a verse before the answer, the answer verses, and a verse after the answer.

   (c) Added the appropriate passage for each record in the dataset.

   I used the same method as Aftab & Malik (2022) to construct the passage for each question and answer. However, the passage construction method in this chapter differs from that in the previous chapter because the QPC file had not yet been published.

Once this was done for all records, they were added to the MRC dataset.

### 6.6.2 Results

The results are divided into two parts: validation and testing. In validation, the dev set of QRCD_v1.2 is used as the test set, while in testing, the test set of QRCD_v1.2 is used.

**Validation:** The QRCD column in Table 6.1 presents the results of the models when they were fine-tuned using only the QRCD dataset. The AraBERT Base model outperformed the other models with a 0.402 pAP@10.

First, I addressed **RQ6.1**, which related to whether further fine-tuning the models using a large CA dataset (QMRC) and the MSA (ARCD) enhanced the performance of the Quran MRC models. The results are shown in the columns 'QRCD AND QMRC,' 'QRCD AND ARCD,' and 'QRCD, QMRC AND ARCD' in Table 6.1 and Table 6.2. There are three interesting observations in the results. First, using the QMRC dataset led to improvements in more than half of the models in terms of pAP@10. The best score was a pAP@10 of 0.482, obtained by AraBERT Large. Eight or seven out of nine models improved on all three other measures. The AraBERT Base model achieved the highest scores on two metrics: 0.881 for F1@1 and 0.794 for pRR, respectively. Second, when I trained the model using the ARCD dataset, it enhanced only the performance of the AraBERT Base model, with a pAP@10 of 0.433, and the MARBERT model, which achieved 0.636 for F1@1 and 0.479 for pRR. Third, using QMRC and ARCD simultaneously to train the AraBERT Base model improved the results to 0.49 for pAP@10, 0.801 for pRR, and 0.885 for F1@1 compared to using QMRC and ARCD separately. I chose the AraBERT Base model alone for fine-tuning on two datasets, QMRC and ARCD. I did not fine-tune the other models because AraBERT Base was the only model that showed improved results when fine-tuned on the two datasets separately. Thus, the answer to the question is that using the CA language dataset (QMRC) in training improved a large number of models across different metrics, while using the MSA dataset (ARCD) or using both datasets together improved only one model.

Based on the above, I answer the following research question: **RQ3:** Does the size of the new Quran question-answer dataset affect the performance of the pre-trained transformer models? Yes, increasing the dataset's size improves the models' performance.

Second, I addressed **RQ6.2**, which relates to whether ensembling the further fine-tuned models and then applying post-processing steps improves the results. The results are shown in Table 6.3. For the first factor of the second question **RQ6.2**, I used the ensemble method for all the models; however, this approach did not yield the best performance, resulting in 0.466 pAP@10. I then ensembled two of the best-performing individual models, which were AraBERT Base and AraBERT Large. The results of this ensemble outperformed the other models, achieving 0.517 pAP@10. The answer to question **RQ6.2** is that the ensemble method for all models did not improve the results, while the ensemble method for the two best models improved only one measure (pAP@10). For the second factor, I noted that the post-processing step improved the results for pAP@10 based on the Ensemble '**POST (Best)**' row shown in Table 6.3.

| **Training Datasets** | **QRCD** | | | | **QRCD AND QMRC** | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **EM** | **F1@1** | **pRR** | **pAP@10** | **EM** | **F1@1** | **pRR** | **pAP@10** |
| AraBERT Large | 0.0 | 0.399 | 0.263 | 0.165 | 0.196 | 0.804 | 0.711 | 0.482 |
| AraBERT Base | 0.110 | 0.799 | 0.680 | 0.402 | 0.159 | 0.881 | 0.794 | 0.458 |
| MARBERT | 0.085 | 0.600 | 0.459 | 0.326 | 0.0 | 0.274 | 0.200 | 0.089 |
| ARBERT | 0.092 | 0.676 | 0.521 | 0.357 | 0.104 | 0.817 | 0.700 | 0.380 |
| QARiB | 0.067 | 0.665 | 0.523 | 0.307 | 0.049 | 0.744 | 0.598 | 0.301 |
| CAMeL-BERT | 0.110 | 0.774 | 0.670 | 0.401 | 0.122 | 0.860 | 0.762 | 0.406 |
| ArabicBERT | 0.079 | 0.712 | 0.556 | 0.332 | 0.104 | 0.754 | 0.617 | 0.330 |
| AraELECTRA | 0.006 | 0.612 | 0.479 | 0.332 | 0.018 | 0.807 | 0.697 | 0.248 |
| CL-AraBERT | 0.110 | 0.784 | 0.648 | 0.373 | 0.116 | 0.799 | 0.714 | 0.383 |

Table 6.1: The validation set result of fine-tuned different Arabic pre-trained models by using different combinations of the datasets part 1. If a model fine-tuned on other datasets performs better than one fine-tuned only on QRCD, it is highlighted in blue.

| Training Datasets | QRCD AND ARCD | | | | QRCD , QMRC AND ARCD | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **EM** | **F1@1** | **pRR** | **pAP@10** | **EM** | **F1@1** | **pRR** | **pAP@10** |
| AraBERT Large | 0.0 | 0.369 | 0.263 | 0.162 | | | | - |
| AraBERT Base | 0.110 | 0.775 | 0.658 | 0.433 | 0.153 | 0.885 | 0.801 | 0.49 |
| MARBERT | 0.073 | 0.636 | 0.479 | 0.291 | | | | - |
| ARBERT | 0.067 | 0.726 | 0.588 | 0.343 | | | | - |
| QARiB | 0.036 | 0.633 | 0.468 | 0.278 | | | | - |
| CAMeL-BERT | 0.067 | 0.737 | 0.628 | 0.362 | | | | - |
| ArabicBERT | 0.055 | 0.681 | 0.559 | 0.313 | | | | - |
| AraELECTRA | 0.006 | 0.555 | 0.431 | 0.218 | | | | - |
| CL-AraBERT | 0.085 | 0.744 | 0.596 | 0.358 | | | | - |

Table 6.2: The validation set result of fine-tuned different Arabic pre-trained models by using different combinations of the datasets part 2. If a model fine-tuned on other datasets performs better than one fine-tuned only on QRCD, it is highlighted in blue. The AraBERT Base model was fine-tuned on the QRCD and QMRC datasets, as well as on the QRCD and ARCD datasets separately. It was then fine-tuned on all three datasets: QRCD, QMRC, and ARCD. The highest score for each metric, compared across these three experiments, is highlighted in red.

| | Best Results | | | |
|---|---|---|---|---|
| **Model** | **EM** | **F1@1** | **pRR** | **pAP@10** |
| AraBERT Large | 0.196 | 0.804 | 0.711 | 0.482 |
| AraBERT Base | 0.153 | 0.885 | 0.801 | 0.49 |
| MARBERT | 0.085 | 0.600 | 0.459 | 0.326 |
| ARBERT | 0.104 | 0.817 | 0.700 | 0.380 |
| QARiB | 0.067 | 0.665 | 0.523 | 0.307 |
| CAMeL-BERT | 0.122 | 0.860 | 0.762 | 0.406 |
| ArabicBERT | 0.079 | 0.712 | 0.556 | 0.332 |
| AraELECTRA | 0.006 | 0.612 | 0.479 | 0.332 |
| CL-AraBERT | 0.116 | 0.799 | 0.714 | 0.383 |
| Ensemble **Vanilla** (All) | 0.190 | 0.784 | 0.627 | 0.466 |
| Ensemble **Vanilla** (Best) | 0.196 | 0.853 | 0.768 | 0.517 |
| Ensemble **POST** (Best) | 0.196 | 0.406 | 0.391 | 0.537 |

Table 6.3: The validation set results of the ensemble approach. The best results for each model are presented first. Ensemble **Vanilla** (All) refers to the ensemble approach to all models. Ensemble **Vanilla** (Best) represents the ensemble approach to the best two performed models (the AraBERT Large and the AraBERT Base). Ensemble **POST** (Best) refers to the **Vanilla** (Best) after applying the post-processing step.

**Testing:** For the test set, I chose two methods based on the performance of the development set. They were (1) the ensemble of AraBERT Base and AraBERT Large with post-processing, which achieved the highest results in all experiments, and (2) the AraBERT Base model, which attained the highest value among the individual models. The ensemble with the post-processing approach achieved a 0.498 pAP@10, while the AraBERT Base model achieved the best performance with a 0.5 pAP@10, as can be seen in Table 6.4.

| **Model** | **EM** | **F1@1** | **pRR** | **pAP@10** |
|---|---|---|---|---|
| Ensemble **POST** (Best) | 0.098 | 0.345 | 0.341 | 0.498 |
| AraBERT Base | 0.095 | 0.738 | 0.630 | 0.5 |

Table 6.4: Test set results.

### 6.6.3 Analysis and Discussion

When I analyzed the answers provided by the models used in the experiments, I identified the following:

1. One of the examined models functioned as a simple match model. When part of the passage contained words from the question, it retrieved this part as an answer, even though the meaning of this portion did not address the question, as shown in Example 3.

   **Example 3:**

   - **"pq_id":** "28:85-88_322":
   - **"question":**

     هل تدبر **القرآن فرض** ؟

     Is meditating on the **Quran ordained**?
   - **"Gold answer":** "[]"
   - **"Predicted answer":**

     إن الذي **فرض** عليك **القرآن** لرادك إلى معاد

     "Verily He Who **ordained** the **Quran** for thee, will bring thee back to the Place of Return."

2. One of the examined models worked as a simple match model. Therefore, the system failed to predict the correct answer when the answer has semantically similar words to the question as shown in Example 4.

   **Example 4:**

   - **"pq_id":** "11:50-60_337"
   - **"question":**

     ما هي الإشارات للدماغ أو لأجزاء من الدماغ في القرآن؟

     What are the references to the brain or parts of the brain in the Quran?

- **"passage":**

وإلى عاد أخاهم هودا قال يا قوم اعبدوا الله ما لكم من إله غيره إن أنتم إلا مفترون . يا قوم لا أسألكم عليه أجرا إن أجري إلا على الذي فطرني أفلا تعقلون . ويا قوم استغفروا ربكم ثم توبوا إليه يرسل السماء عليكم مدرارا ويزدكم قوة إلى قوتكم ولا تتولوا مجرمين . قالوا يا هود ما جئتنا ببينة وما نحن بتاركي آلهتنا عن قولك وما نحن لك بمؤمنين . إن نقول إلا اعتراك بعض آلهتنا بسوء قال إني أشهد الله واشهدوا أني بريء مما تشركون . من دونه فكيدوني جميعا ثم لا تنظرون . إني توكلت على الله ربي وربكم ما من دابة إلا هو آخذ بناصيتها إن ربي على صراط مستقيم . فإن تولوا فقد أبلغتكم ما أرسلت به إليكم ويستخلف ربي قوما غيركم ولا تضرونه شيئا إن ربي على كل شيء حفيظ . ولما جاء أمرنا نجينا هودا والذين آمنوا معه برحمة منا ونجيناهم من عذاب غليظ . وتلك عاد جحدوا بآيات ربهم وعصوا رسله واتبعوا أمر كل جبار عنيد . وأتبعوا في هذه الدنيا لعنة ويوم القيامة ألا إن عادا كفروا ربهم ألا بعدا لعاد قوم هود

"To the 'Ad People (We sent) Hud, one of their own brethren. He said: "O my people! worship Allah! ye have no other god but Him. (Your other gods) ye do nothing but invent! (50) "O my people! I ask of you no reward for this (Message). My reward is from none but Him who created me: Will ye not then understand? (51) "And O my people! Ask forgiveness of your Lord, and turn to Him (in repentance): He will send you the skies pouring abundant rain, and add strength to your strength: so turn ye not back in sin!" (52) They said: "O Hud! No Clear (Sign) that hast thou brought us, and we are not the ones to desert our gods on thy word! Nor shall we believe in thee! (53) "We say nothing but that (perhaps) some of our gods may have seized thee with imbecility." He said: "I call Allah to witness, and do ye bear witness, that I am free from the sin of ascribing, to Him, (54) "Other gods as partners! so scheme (your worst) against me, all of you, and give me no respite. (55) "I put my trust in Allah, My Lord and your Lord! There is not a moving creature, but He hath grasp of its fore-lock. Verily, it is my Lord that is on a straight Path. (56) "If ye turn away,- I (at least) have conveyed the Message with which I was sent to you. My Lord will make another people to succeed you, and you will not harm Him in the least. For my Lord hath care and watch over all things." (57) So when Our decree issued, We saved Hud and those who believed with him, by (special) Grace from Ourselves: We saved them from a severe penalty. (58) Such were the 'Ad People: they rejected the Signs of their Lord and Cherisher; disobeyed His messengers; And followed the command of every powerful, obstinate transgressor. (59) And they were pursued by a Curse in this life,- and on the Day of Judgment. Ah! Behold! for the 'Ad rejected their Lord and Cherisher! Ah! Behold! removed (from sight) were 'Ad the people of Hud!"

- **"Gold answer":**

  ناصيتها

  "its fore-lock"

  (a) **"Predicted answer":**

  يرسل السماء عليكم مدرارا ويزدكم قوة إلى قوتكم

  "He will send you the skies pouring abundant rain, and add strength to your strength"

  **"rank":** 1, ...

  (b) **"Predicted answer":**

  ولا تتولوا مجرمين . قالوا يا هود ما جئتنا ببينة وما نحن بتاركي آلهتنا عن قولك وما نحن لك بمؤمنين

  "so turn ye not back in sin!" (52) They said: "O Hud! No Clear (Sign) that hast thou brought us, and we are not the ones to desert our gods on thy word! Nor shall we believe in thee!"

  **"rank":** 2 ...

(c) **"Predicted answer":**

إن نقول إلا اعتراك بعض آلهتنا بسوء قال إني أشهد الله واشهدوا أني
"We say nothing but that (perhaps) some of our gods may have seized thee with imbecility." He said: "I call Allah to witness, and do ye bear witness, that I am"
**"rank":** 3 ...

(d) **"Predicted answer":**

استغفروا ربكم ثم توبوا إليه
"And O my people! Ask forgiveness of your Lord, and turn to Him (in repentance"
**"rank":** 4 ...

The conclusion I reached from this analysis is that the 'understanding' (or perhaps a deeper encoding) of these models regarding this type of text requires a great deal of further development and research to reach the level of human understanding.

## 6.7 Experiment 2: Answering Questions from the Hadith

### 6.7.1 Datasets

In this study, I used three different datasets, as follows:

1. **ARCD:** Its details are explained in Section 5.6.1.

2. **QMRC:** Its details are explained in Section 6.6.1.

3. **Hadith MRC (HMRC):** It is an MRC dataset. HMRC is a question-passage-answer triplet dataset, while HAQA is a question-answer pairs dataset. HMRC is an extended version of the HAQA dataset, created by adding a paragraph for each record. Hadith books are divided into sections,

and each section deals with a specific topic. Therefore, I collected four consecutive hadiths from each book into one longer text and numbered each passage. Then, I added the passage that contains the hadith found in the answer to each record. The resulting corpus consisted of 1,359 questions and 1,598 records. Its distribution is shown in Table 6.5. An example of a record from the HMRC in Arabic is shown in Example 5.

**Example 5:**

- **Question ID:** 96
- **Question:**

بماذا نحلف ؟

What do we swear to?

- **The Paragraph:**

... أن النبي صلى الله عليه وسلم قال من كان حالفا فليحلف بالله أو ليصمت.
سمع رسول الله صلى الله عليه وسلم يقول: إذا سمعتم النداء فقولوا مثل ما يقول ...

... Who must take an oath, may do so by swearing in the Name of Allah or he should remain silent. I heard the Messenger of Allah saying, "When you hear the Adhan, repeat what the Mu'adhdhin says . . . .

- **The Answer:**

بالله

By God

### 6.7.2 Results

In this study, I used four evaluation metrics: EM, F1, pRR, and pAP@10 as part of Experiment 1 in this chapter. Each of these four metrics was computed for every question. I then calculated the overall score for each metric by averaging it across all questions. The results of the experiments on the use of pre-trained

| Dataset | Question-Passage-Answer Triplets | Percentage |
|---|---|---|
| Development | 160 | 10% |
| Training | 1,038 | 65% |
| Test | 400 | 25% |
| All | 1,598 | 100% |

Table 6.5: The distribution of the HMRC.

transformer-based models in the MRC task to answer Hadith questions are shown in Table 6.6, Table 6.7, and Table 6.8.

First, I began by addressing **RQ6.4**, which was related to studying the effects of fine-tuning the models using larger CA datasets (such as QMRC) or/and MSA datasets (such as ARCD) on their performance. The results are shown in the columns under "HMRC+ARCD", "HMRC+ARCD+QMRC" and "HMRC+QMRC" in Table 6.6 and Table 6.7. The answer to this research question is divided into three parts.

- First, training the models with the ARCD led to improvements in all four evaluation metrics across four models: AraBERT Large, AraBERT Base, MARBERT, and CAMeL-BERT. The EM and pRR were the most improved evaluation metrics, with eight out of nine models showing improvement in these two. The best scores were an F1@10 of 0.700 and a pRR of 0.643, obtained by CAMeL-BERT. The AraBERT Large model achieved the highest result in EM with 0.239, while the AraBERT Base model achieved the highest result in pAP@10 with 0.229.

- Second, using the QMRC dataset in the training phase enhanced only AraBERT Large regarding the four metrics. When the QMRC was used, six of the nine models improved their EM scores, three improved their F1@10 scores, and two improved their pAP@10 and pRR scores. With this training dataset, the highest EM score was 0.209, the F1@10 score was 0.668, the pRR score was 0.589, and the pAP@10 score was 0.224, achieved by the AraBERT Large, CL-AraBERT, CAMeL-BERT, and AraBERT Base models, respectively.

- Third, only the EM improved when fine-tuning the AraBERT Large model using the ARCD and QMRC datasets. The EM score was 0.242, compared to 0.209 when fine-tuning with QMRC and 0.239 when fine-tuning with ARCD, as shown in Table 6.6 and Table 6.7. The other metrics—F1@10, pRR, and pAP@10—were 0.647, 0.578, and 0.213, respectively. These scores were worse than when training the model using only ARCD but better than when training the model using only QMRC. The F1@10, pRR, and pAP@10 scores for training the model with only ARCD were 0.663, 0.588, and 0.218, respectively, while the scores for training the model with only QMRC were 0.642, 0.560, and 0.204, respectively. I chose the AraBERT Large model because it obtained improved results when trained individually using the ARCD and QMRC.

So, the answer to the question **RQ6.4** is that using the MSA dataset in the training substantially improved the results, while using the CA dataset or both datasets slightly improved the results.

| Training Datasets | HMRC | | | | HMRC and ARCD | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **EM** | **F1@10** | **pRR** | **pAP@10** | **EM** | **F1@10** | **pRR** | **pAP@10** |
| AraBERT Large | 0 | 0.158 | 0.099 | 0.051 | 0.239 | 0.663 | 0.588 | 0.218 |
| AraBERT Base | 0.118 | 0.67 | 0.604 | 0.227 | 0.215 | 0.676 | 0.612 | 0.229 |
| MARBERT | 0.015 | 0.509 | 0.329 | 0.154 | 0.054 | 0.595 | 0.454 | 0.186 |
| ARBERT | 0.072 | 0.645 | 0.556 | 0.215 | 0.16 | 0.661 | 0.579 | 0.214 |
| QARiB | 0.009 | 0.381 | 0.224 | 0.104 | 0.009 | 0.412 | 0.235 | 0.107 |
| CAMeL-BERT | 0.106 | 0.68 | 0.604 | 0.226 | 0.181 | 0.700 | 0.643 | 0.237 |
| ArabicBERT | 0.093 | 0.598 | 0.504 | 0.199 | 0.121 | 0.596 | 0.507 | 0.198 |
| AraELECTRA | 0 | 0.408 | 0.301 | 0.133 | 0.006 | 0.342 | 0.228 | 0.106 |
| CL-AraBERT | 0.13 | 0.68 | 0.597 | 0.225 | 0.154 | 0.673 | 0.599 | 0.222 |

Table 6.6: Results of using different Arabic MRC datasets to fine-tune different Arabic pre-trained transformer-based models part 1. If a model fine-tuned on other datasets performs better than one fine-tuned only on HMRC, it is highlighted in blue.

| Training Datasets | HMRC and QMRC | | | | HMRC , ARCD and QMRC | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **EM** | **F1@10** | **pRR** | **pAP@10** | **EM** | **F1@10** | **pRR** | **pAP@10** |
| AraBERT Large | 0.209 | 0.642 | 0.560 | 0.204 | 0.242 | 0.647 | 0.578 | 0.213 |
| AraBERT Base | 0.178 | 0.662 | 0.584 | 0.224 | | | | |
| MARBERT | 0.012 | 0.52 | 0.325 | 0.154 | | | | |
| ARBERT | 0.1 | 0.613 | 0.516 | 0.199 | | | | |
| QARiB | 0.009 | 0.368 | 0.207 | 0.100 | | | | |
| CAMeL-BERT | 0.166 | 0.662 | 0.589 | 0.221 | | | | |
| ArabicBERT | 0.106 | 0.579 | 0.49 | 0.189 | | | | |
| AraELECTRA | 0 | 0.613 | 0.551 | 0.209 | | | | |
| CL-AraBERT | 0.145 | 0.668 | 0.587 | 0.212 | | | | |

Table 6.7: Results of using different Arabic MRC datasets to fine-tune different Arabic pre-trained transformer-based models part 2. If a model fine-tuned on other datasets performs better than one fine-tuned only on HMRC, it is highlighted in blue. The AraBERT large model was fine-tuned on the HMRC and ARCD datasets, as well as on the HMRC and QMRC datasets separately. It was then fine-tuned on all three datasets: HMRC, ARCD, and QMRC. The highest score for each metric, compared across these three experiments, is highlighted in red.

**RQ6.5** was related to whether the ensemble approach, followed by applying post-processing steps, enhanced the performance of the Hadith MRC models. This question was answered by selecting the best-performing version of each model, applying the ensemble approach to them, and comparing the results, as shown in Table 6.8. Unfortunately, the results of the ensemble approach performed worse than the CAMeL-BERT model in terms of F1@10, pRR, and pAP@10, as indicated in Table 6.8. The ensemble approach achieved a score of 0.597 for F1@10, 0.503 for pRR, and 0.198 for pAP@10, while the best model alone (CAMeL-BERT) achieved scores of 0.700 for F1@10, 0.643 for pRR, and 0.237 for pAP@10. Additionally, AraBERT Large achieved 0.239 for EM compared to 0.221 for EM, which was obtained by the ensemble approach. I noticed that the poor performance of some models negatively affected the performance of the ensemble approach. Therefore, I applied the ensemble approach only to the best models: CAMeL-BERT Base and CL-AraBERT. The results of this ensemble showed improvement only in EM, with a score of 0.251.

In summary, the answer to the question **RQ6.5** is that the ensemble approach improved the results when the individual results of the included models were close. However, if there was a significant contrast in the individual results of the models, then the poorer-performing models negatively affected the results of the ensemble approach. Using the post-processing steps yielded very poor results in general, as shown in Table 6.8.

To address **RQ6.3**, which is related to how well the pre-trained models performed in the MRC task of answering questions from the Hadith, I conducted a number of experiments. The results are shown in Table 6.6, Table 6.7 and Table 6.8. The results showed that the models performed well in general. Five models achieved the best results when they trained with HMRC and ARCD, three models achieved the best results when they trained with HMRC only, and one model achieved the best results when they trained with QMRC. The best models were CAMeL-BERT with an F1@10 score of 0.700, pRR score of 0.643 and pAP@10 score of 0.237.

| | Best Results | | | |
|---|---|---|---|---|
| **Model** | **EM** | **F1@10** | **pRR** | **pAP@10** |
| AraBERT Large | 0.239 | 0.663 | 0.588 | 0.218 |
| AraBERT Base | 0.215 | 0.676 | 0.612 | 0.229 |
| MARBERT | 0.054 | 0.595 | 0.454 | 0.186 |
| ARBERT | 0.072 | 0.645 | 0.556 | 0.215 |
| QARiB | 0.009 | 0.412 | 0.235 | 0.107 |
| CAMeL-BERT | 0.181 | 0.700 | 0.643 | 0.237 |
| ArabicBERT | 0.093 | 0.598 | 0.504 | 0.199 |
| AraELECTRA | 0 | 0.613 | 0.551 | 0.209 |
| CL-AraBERT | 0.13 | 0.68 | 0.597 | 0.225 |
| Ensemble **Vanilla** (All) | 0.221 | 0.597 | 0.503 | 0.198 |
| Ensemble **Vanilla** (Best) | 0.251 | 0.698 | 0.633 | 0.225 |
| Ensemble **POST** (Best) | 0.242 | 0.194 | 0.190 | 0.184 |

Table 6.8: The results of the ensemble approach. The best results for each model are presented first. Ensemble **Vanilla** (All) refers to the ensemble approach to all models. Ensemble **Vanilla** (Best) represents the ensemble approach to the best two performed models (the CAMeL-BERT model and AraBERT Base). Ensemble **POST** (Best) refers to the **Vanilla** (Best) after applying the post-processing step.

### 6.7.3 Analysis and Discussion

I analysed the expected answers to determine why the models failed to predict the correct answer in some cases. I noted the following points:

1. The models failed to answer a number of questions completely. It seemed that the models were not always able to extract the answer based on the type of question. For example, a question that begins with "who" should be answered by a noun; however, the model often extracted stop words or a description, as illustrated in Example 6, or pronouns, as shown in Example 7, as demonstrated below:

   **Example 6:**

   - **Question ID:** 120
   - **Question:**

     **من** هما الملكان الموكلان بالسؤال في القبر؟

     **Who** are the two angels entrusted with the question in the grave?
   - **The Passage:**

     ... قال رسول الله صلى الله عليه وسلم إذا قبر الميت أو قال أحدكم أتاه ملكان أسودان أزرقان يقال لأحدهما المنكر وللآخر النكير...

     ...When the dead - or one of you - is buried, two dark and blue angels will come to him; one is called 'Munkir' and the other is called 'Nakir'...
   - **The Gold answer:**

     يقال لأحدهما المنكر وللآخر النكير

     One is called 'Munkir', and the other is called 'Nakir'.
   - **The Predicted answer:**

     **أتاه ملكان أسودان أزرقان**

     **Two dark and blue angels will come to him**

**Example 7:**

- **Question ID:** 365
- **Question:**

  من هو الذي ساقه في الميزان أثقل من جبل أحد ؟

  **Who** is the one whose leg in the balance is heavier than the mountain of Uhud?
- **The Passage:**

  ... أمر النبي صلى الله عليه وسلم ابن مسعود فصعد على شجرة أمره أن يأتيه منها بشيء فنظر أصحابه إلى ساق عبد الله بن مسعود حين صعد الشجرة فضحكوا من حموشة ساقيه فقال رسول الله صلى الله عليه وسلم ما تضحكون لرجل عبد الله أثقل في الميزان يوم القيامة من أحد..

  ...The Prophet Peace be upon him instructed lbn Mas'ood to climb up a tree and he told him to bring him something from it, and his companions looked at the shins of Abdullah bin Mas'ood when he climbed the tree and laughed at how thin his shins were, the Messenger of Allah Peace be upon him said: "Why are you laughing? The leg of Abdullah will be heavier in the balance on the Day of Resurrection than (Mount) Uhud....
- **The Gold answer:**

  ابن مسعود

  lbn Mas'ood
- **The Predicted answer:**

  **منها**

  **From it**

2. Example 8 illustrates one of the examined model's difficulties in answering a question. The reason may be that the terms in the passage are synonymous with those in the question.

**Example 8:**

- **Question ID:** 132
- **Question:**

ما هو الشيء الذي **يهرب منه الشيطان** ولا يطيق سماعه ؟

What is the thing that **evil runs away** from and cannot bear to hear?

- **The Passage:**

... أن رسول الله صلى الله عليه وسلم قال إذا نودي للصلاة **أدبر الشيطان** ...

...God's Messenger says, "When a summons to prayer is made **the evil turns his back**"...

- **The Gold answer:**

إذا نودي للصلاة

When a summons to prayer is made

- **The Predicted answer:**

-

3. One of the examined models does not deeply "understand" the question to predict the answer correctly; instead, it applies a direct and simple matching method, which leads to an incorrect answer, as illustrated in Example 9.

**Example 9:**

- **Question ID:** 238
- **Question:**

ما هي **الزيادة** في قوله تعالى **للذين أحسنوا الحسنى وزيادة**؟

What is the meaning of **more** in saying of Allah Most High: And **for those who have done good is the best and even more**?

- **The Passage:**

  ... عن النبي صلى الله عليه وسلم في قول الله عزوجل **للذين أحسنوا الحسنى وزيادة** قال إذا دخل أهل الجنة الجنة نادى مناد إن لكم عند الله موعدا يريد أن ينجزكموه قالوا ألم تبيض وجوهنا وتنجنا من النار وتدخلنا الجنة قال فيكشف الحجاب قال فوالله ما أعطاهم الله شيئا أحب إليهم من النظر إليه...

  ...from the Prophet peace be upon him, regarding the saying of Allah Most High: And **for those who have done good is the best and even more (10:26)** - He Peace be upon him said: "When the inhabitants of Paradise have entered Paradise a caller will call out: 'Indeed there remains for you a promise with Allah, and He wants to reward you with it.' They will say: 'Have your faces not been made bright, have we not been saved from the Fire, and have we not been admitted into Paradise?'" He said: "So the Veil will be lifted." He said: "By Allah! Nothing given to them [by Allah] will be more beloved to them than looking at Him....."

- **The Gold answer:**

  النظر إليه

  looking at Him

- **The Predicted answer:**

  للذين أحسنوا الحسنى وزيادة

  for those who have done good is the best and even more

4. Example 10 illustrates one of the examined model's failures to accurately 'understand' the passage, as it was unable to extract the noun to which the pronoun refers.

   **Example 10:**

   - **Question ID:** 204

- **Question:**

  من هو أول من تنشق عنه الأرض ؟

  Who is the first one from whom the earth will open?

- **The Passage:**

  ... عن النبي صلى الله عليه وسلم قال أنا سيد ولد آدم و أول من تنشق عنه الأرض..

  ...Messenger of Allah (May peace be upon him) as saying: I shall be pre-eminent among the descendants of Adam, the first from whom the earth will be cleft open ...

- **The Gold answer:**

  النبي صلى الله عليه وسلم

  Messenger of Allah

- **The Predicted answer:**

  أنا

  I

The results of the analysis indicate that these models need significant improvements and further research to enhance their comprehension or deeper representation of this text type in order to approach human-level understanding.

## 6.8 Answering the Research Questions

Based on the experiments conducted in this chapter, the following question can be answered: **RQ2:** Can PLMs answer questions from the Quran and Hadith in Arabic? The results presented in Table 6.1, Table 6.2, Table 6.3, Table 6.6, Table 6.7, and Table 6.8 demonstrate acceptable model performance but highlight the need for further improvement. Our findings may pave the way for additional research on MRC tasks in Arabic and other languages.

## 6.9 Conclusion

One of the goals of this thesis is to build a model to answer questions in the Holy Quran and Hadith using transformer-based models. The model includes two tasks, one of which is the MRC task. This chapter presents a literature survey on the existing models that show good performance. It also discusses existing and potential improvement methods. Models were built to answer questions from the Quran and Hadith and implement improvement methods.

The proposed MRC system combines transfer learning and ensemble approaches for the best-performing models. Initially, I fine-tuned nine different Arabic pre-trained models using different data collections. I then applied the ensemble approach to the two best-performing models. Finally, I implemented appropriate post-processing steps.

In the Quran, the ensemble of the base and large variants of AraBERT achieved the best results on the development set, with a 0.537 pAP@10. The second-highest score was achieved by base AraBERT with a 0.49 pAP@10. The results of applying these two models to the test set showed that the base AraBERT model was the best with a score of 0.5 pAP@10, while the ensemble model achieved a score of 0.49 pAP@10.

In the Hadith, the best models were CAMeL-BERT with an F1@10 score of 0.700, a pRR score of 0.643, and a pAP@10 score of 0.237.

# Chapter 7

# Evaluating the Large Language Models (LLMs) (GPT-4) as Questions Answering Model

## 7.1 Introduction

This chapter investigates GPT-4 as a QA model for the Holy Quran and Hadith. Firstly, the chapter presents a literature survey of studies that used GPT as a model to answer questions in different fields and languages. Then, the chapter describes two experiments: the first answers questions of various types related to the Holy Quran, and the second answers questions of various types related to the Prophet's Hadiths. Following this, the results are displayed. Finally, the results are analyzed, with some examples of errors discussed. Part of this chapter is published in Alnefaie *et al.* (2023b).

## 7.2 Related Work

Recently, OpenAI has developed and provided access to several versions of the LLM-driven ChatGPT—"GPT" refers to its underlying generative pre-trained transformer model (Brown *et al.*, 2020; Ouyang *et al.*, 2022). Substantial progress has been achieved in NLP by developing LLMs such as GPT. These models undergo extensive training on large text datasets. This training enables them to produce text that closely resembles human-generated content, provide accurate responses to inquiries, and proficiently handle various NLP tasks (Kasneci *et al.*, 2023).

Several studies have focused on testing GPT on downstream tasks (Jiao *et al.*, 2023; Qin *et al.*, 2023; Wang *et al.*, 2023b). Katz *et al.* (2023) conducted an experiment to study the performance of GPT in passing the Uniform Bar Examination (UBE). Success on this exam is a condition for law practice in most states within the US. It consists of several components, including multiple-choice and essay items. GPT-4 (i.e., the fourth version of ChatGPT's underlying model) achieved good results with 297 points. Kung *et al.* (2023) suggested measuring the performance of GPT-3.5 in the United States Medical Licensing Exam (USMLE). The results showed that it was close to passing the test. Wood *et al.* (2023) recommend investigating the efficiency of GPT-3.5 using more than 28,000 questions from accounting tests. They found that the average results of the model were better than students when partially correct answers were considered. Kasai *et al.* (2023) proposed studying the performance of various versions of GPT using the Japanese national medical licensing examinations from the last six years. The results showed that GPT-4 had the best performance and passed all exams. However, to the best of our knowledge, there has not yet been a published examination of how well GPT performs in answering Islamic questions.

Our research sought to answer the question, **RQ5:** "Is GPT a good Islamic expert?" To address this question, GPT's performance was tested with a set of questions to assess its behavior, strengths, and weaknesses. The results of this study will benefit a large segment of Muslims worldwide, motivate further research to address any identified defects, and assist researchers in choosing an appropriate language model in the future.

## 7.3 Proposed Model

The development of GPT has progressed through several generations: GPT-1 (Radford *et al.*, 2018), GPT-2 (Radford *et al.*, 2019), GPT-3 (Brown *et al.*, 2020), GPT-3.5 (Ye *et al.*, 2023), and GPT-4 (OpenAI, 2023). Various techniques have been employed in these different versions, which have positively impacted the performance and development of the language model.

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

## 7.4 Experiment 1: Answering Questions from the Holy Quran

### 7.4.1 Datasets

QUQA was used in this experiment. It contains three kinds of questions: confirmation (193 questions), descriptive (1,752 questions), and fact (241 questions). The total number of questions is 2,186.

### 7.4.2 Evaluation

In response to the supplied questions, the output generated by GPT-4 contains natural-sounding text and several series of Quranic verses, as shown in Example 1. In the output, the series may consist of one or more verses. Therefore, I evaluated the textual answers and the Quranic verses independently.

**Example 1:**

- **Question ID:** 522
- **Question:**

  أجب عن السؤال التالي من القرآن الكريم ، مع ذكر الآية التي تجيب على السؤال:.
  ما عدد السموات؟

  Answer the following question from the Holy Quran, citing the verse that answers it: How many heavens are there?

- **The GPT Answer:**

  **عدد السموات ذُكِرَ في القرآن الكريم بأنها سبع سماوات.** يمكن العثور على هذا الرقم في عدة آيات مثل: ١) سورة البقرة (الآية ٢٩): آهُوَ الَّذِي خَلَقَ لَكُم مَّا فِي الْأَرْضِ جَمِيعًا ثُمَّ اسْتَوَىٰ إِلَى السَّمَاءِ فَسَوَّاهُنَّ سَبْعَ سَمَاوَاتٍ وَهُوَ بِكُلِّ شَيْءٍ عَلِيمٌ. ٢) سورة الإسراء (الآية ٤٤): آتُسَبِّحُ لَهُ السَّمَاوَاتُ السَّبْعُ وَالْأَرْضُ وَمَن فِيهِنَّ وَإِن مِّن شَيْءٍ إِلَّا يُسَبِّحُ بِحَمْدِهِ وَلَـكِن لَّا تَفْقَهُونَ تَسْبِيحَهُمْ إِنَّهُ كَانَ حَلِيمًا غَفُورًا

  **The number of heavens mentioned in the Holy Quran is seven.** This number can be found in several verses, such as (1) Surah Al-Baqarah (verse 29), “He it is Who created for you all that is on Earth. Then He Istawa (rose over) towards the heaven and made them seven heavens and He is the All-Knower of everything"; (2) Surah Al-Isra (verse 44), “The seven heavens and the Earth and whatever is in them exalt Him. And there is not a thing except that it exalts [ Allah ] by His praise, but you do not understand their [way of] exalting. Indeed, He is ever Forbearing and Forgiving."

The evaluation of the Quranic verses was both automatic and manual. First, I manually extracted the Quranic verses from the text output by GPT-4 because they were in different formats and put them in a standardised form. I wrote a programme to check whether the text of a Quranic verse was fake and then validated it as an answer to the question using the ’golden’ labeled dataset. I noticed that some of the verses mentioned in the GPT answer correctly addressed the question but were not mentioned in the golden dataset. Therefore, I checked all the answers and added unanticipated ’found’ correct answers to the dataset, as appropriate (only during the evaluation step). For example, all the verses cited in the answer given in Example 1 are correct, but only the first verse was originally mentioned in the correct answer in the dataset. I evaluated the answers manually because the GPT answers may have a similar meaning to the golden answers but use different words. For example, in Question 2117 (’How long is full breastfeeding?’), the answer is ’حولين’, but GPT answered with ’سنتين’; these two words are synonymous in meaning two years. The gold answer was taken from the 'Exper_Commentary' fields in the QUQA dataset, as shown in Table 4.7.

I considered different retrieved series of Quranic verses in the answer as a ranked list and used F1@1, EM, F1, and pAP as evaluation metrics. F1@1 and EM are typically applied only to the top predicted answer, which, in my case, was the first series. EM was a binary value assigned a value of 1 when the first series of verses matched exactly with one of the gold verses, and 0 if not. To compute F1, I measured the overlap between each series of verses and the golden answer and then took the average. The gold answer was taken from the 'Chapter_Number,' 'Verses_Number_Start,' and 'Verses_Number_End' fields in the QUQA dataset as shown in Table 4.7 and Table 4.8. If F1 was computed only for the first series, it was referred to as F1@1. The pAP measure was used to consider the rank of the correct answers in the retrieved list. If the system could retrieve the correct answers at the top of the list, then the score was higher. The above measures were computed for each question, and then the average was taken. Malhas & Elsayed (2020, 2022) used these four measures to assess their Quran QA system. I treated the text as a bag of words and used EM and F1 to assess it. If the first sentence contained the exact answer, then the EM was assigned a value of 1; otherwise, it was assigned a value of 0. The GPT-4 answers to all the questions are publicly available.[1]

### 7.4.3 Results

This section presents the performance of GPT-4 with the QUQA dataset. I conducted three experiments, each focusing on a particular type of question. The QUQA contains three kinds of questions: confirmation (193 questions), descriptive (1,752 questions), and fact (241 questions). The summary of results for the Quranic series portion of the GPT-4 answers is shown in Table 7.1, while the results for the produced-text portion are shown in Table 7.2.

In this study, I used four evaluation metrics: EM, pAP, F1, and F1@1 scores for these experiments (Malhas & Elsayed, 2020, 2022). When evaluating the natural text portion of the answer, I treated it as a bag of words, similar to the

[1] https://github.com/scsaln/GPT4

method used in Chapter 6. Since all the Quranic verses are presented in full in the answer, I evaluated that part of the Quranic verses as a bag of verses, as follows:

- **EM:**
  EM is a binary score (0 or 1). If one of the gold answers fully matches the top candidate's answer, then the EM equals 1.

- **F1:**
  Let $R$ be the system's retrieved ranked list of answers to the question: $R = [r_1, r_2, r_3, r_4, r_5]$

  Let $A$ be the set of gold direct answers to the question: $A = [a_1, a_2, a_3, a_4, a_5, a_6]$. A is the series of verses found in the gold standard answer across multiple records in QUQA with the same question ID.

  To calculate the value of F1, I need to calculate the following values:

  1. **For each question:**

     (a) **Compute the $\boldsymbol{m_r}$ for each system answer:** I define the answer-matching score ( m ) for a system's response ( r ), represented as $m_r$, as the highest matching score of ( r ) compared to all direct gold standard answers ( A ) for the given question.

$$m_r = max_{a \epsilon A} f_m(r, a) \tag{7.1}$$

where $f_m(r, a)$ is the F1 measure of the system's answer r and the gold answer a.

$$m_r = max(F1(r_1|a_1), F1(r_1|a_2), F1(r_1|a_3), F1(r_1|a_4), F1(r_1|a_5), F1(r_1|a_6)) \tag{7.2}$$

$$F1(r_i|a_j) = \frac{2 * Precision(r_i a_j) * Recall(r_i a_j)}{Precision(r_i a_j) + Recall(r_i a_j)} \tag{7.3}$$

where $i$ from 1 to 5 and $j$ from 1 to 6.

$$Precision(r_i a_j) = \frac{\text{retrieved and relevant verses}}{\text{all retrieved verses}} \tag{7.4}$$

$$Recall(r_i a_j) = \frac{\text{retrieved and relevant verses}}{\text{all relevant verses}} \tag{7.5}$$

For example, Let R be the system's retrieved ranked list of answers to the question.
R=[$r_1$ = 5:3-5]
Let A be the set of gold standard direct answers to the question.
A=[$a_1$=2:1-1, $a_2$=4:4-5, $a_3$=100:4-6, $a_4$=5:3-4]
Chapter_id:Verse_id1-Verse_id2 means the answer is from the verse number id1 to id2 from chapter id in the Quran.
To calculate the value of F1:
$m_r = max(F1(r_1|a_1), F1(r_1|a_2), F1(r_1|a_3), F1(r_1|a_4))$

$$F1(r_1|a_j) = \frac{2 * Precision(r_1 a_j) * Recall(r_1 a_j)}{Precision(r_1 a_j) + Recall(r_1 a_j)} \tag{7.6}$$

where $j$ from 1 to 4.
1-Compute $F1(r_1|a_1)$:
$r_1$ = 5:3-5
$a_1$= 2:1-1
No match.
$F1(r_1|a_1) = 0$
2-Compute $F1(r_1|a_2)$:
$r_1$ = 5:3-5
$a_2$= 4:4-5
No match.
$F1(r_1|a_2) = 0$
3-Compute $F1(r_1|a_3)$:
$r_1$ = 5:3-5
$a_3$= 100:4-6
No match.
$F1(r_1|a_3) = 0$

4-Compute $F1(r_1|a_4)$:
$r_1$ = 5:3-5
$a_4$= 5:3-4
Partial match.

$$Precision(r_1 a_4) = \frac{2}{3} = 0.666 \tag{7.7}$$

$$Recall(r_1 a_4) = \frac{2}{2} = 1 \tag{7.8}$$

$$F1(r_1|a_1) = \frac{2 * 0.666 * 1}{0.666 + 1} = \frac{1.332}{1.666} = 0.799 \tag{7.9}$$

$m_{r_1} = max(0, 0, 0, 0.799)$= 0.799

(b) **Compute the overall $F1$ for the question:**

$$F1 = \frac{\sum_{i=1}^{|R|} m_{r_i}}{|R|} \tag{7.10}$$

In the above example |R|=5, where |R| is the number of answers in the ranked list.

2. **Compute the $F1$ for all questions:**

$$F1 = \frac{\sum_{w=1}^{Q} F1_w}{Q} \tag{7.11}$$

Where $Q$ is the number of questions in the test set.

- **F1@1:**

$$F1@1 = m_{r_1} \tag{7.12}$$

F1@1 is computed only using the first system answer $r_1$.

- **pAP:**

$$pAP = \frac{1}{|A|} \sum_{k=1}^{|R|} 1\{m_{r_k} > 0\} pPrec@k(R) \tag{7.13}$$

Where the number of answers in the ranked list is represented by |R|, and the number of gold answers is represented by |A|. $1\{m_{r_k} > 0\}$ is a function that is set to 1 if the predicted answer at position (k) fully or partially matches one of the gold answers, and set to 0 if it does not match either fully or partially. Partial precision at rank (k) is computed as follows:

$$pPrec@k(R) = \frac{1}{k} \sum_{i=1}^{k} m_{r_i} \tag{7.14}$$

where ( R ) represents the ranked list of predicted answers returned by the system, ( $r_i$ ) denotes the predicted answer at a position or rank ($i$) in ( $R$), and ($m_{r_i}$) refers to the partial matching score of ( $r_i$), as shown in Eq. 7.1.

For example, let R1 be the system's first retrieved ranked list of answers to the question.
R1=[$r_1$=2:1-1, $r_2$=4:4-5, $r_3$=10:40-50]
Let R2 be the system's second retrieved ranked list of answers to the question.
R2=[$r_1$ = 5:3-5, $r_2$=2:1-1, $r_3$=4:4-5,]
Let A be the set of gold standard direct answers to the question.
A=[$a_1$=2:1-1, $a_2$=4:4-5]

1. To compute the value of $pAP$ for the first system:

   - **Find all $m_{r_k}$**
     $m_{r_1}$=1
     $m_{r_2}$=1
     $m_{r_3}$=0
   - **Calculate the value of $pAP$ using Eq. 7.16 and Eq. 7.14.**

$$pAP = \frac{1}{2}\sum_{K=1}^{3} 1\{m_{r_k} > 0\} pPrec@k(R) = \frac{1}{2}(\frac{1}{1} + \frac{1+1}{2} + 0) = \frac{1}{2}(1+1+0) = \frac{1}{2}(2) = 1 \tag{7.15}$$

2. To compute the value of $pAP$ for the second system:

- **Find all $m_{r_k}$**

  $m_{r_1}$=0

  $m_{r_2}$=1

  $m_{r_3}$=1

- **Calculate the value of $pAP$ using Eq. 7.16 and Eq. 7.14.**

$$pAP = \frac{1}{2}\sum_{K=1}^{3} 1\{m_{r_k} > 0\} pPrec@k(R) = \frac{1}{2}(0 + \frac{0+1}{2} + \frac{0+1+1}{3}) = \frac{1}{2}(\frac{1}{2} + \frac{2}{3}) = \frac{1}{2}(0.5 + 0.666) = \frac{1}{2}(1.166) = 0.83 \tag{7.16}$$

pAP of System 1 = 1

pAP of System 2 = 0.83

The pAP value in the first system is higher than that in the second system. This is because the gold answers in the first system appeared at higher ranks compared to those in the second system.

In general, GPT-4 did not achieve impressive results in answering Islamic questions, as claimed in the state of the art, when answering questions from other domains. The results are shown in Table 7.1 and Table 7.2. The sentences and the series of Quranic verses in their answers to fact-type questions outperformed the other types, with a 0.3 F1@1 score, 0.27 pAP, and a 0.25 EM score for the Quran verse portion, and a 0.34 EM score for the regular text. For the entire regular text in the answer, GPT-4 achieved higher results for confirmation-type questions, with a 0.29 F1 score. Additionally, GPT-4 obtained a high score of 0.36 F1 for the entire series of Quranic verses when answering descriptive questions.

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

Table 7.1: The evaluation results of the Quranic series portion of the GPT-4 answers.

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

Table 7.2: The evaluation results of the produced-text portion of the GPT-4 answers.

The ranking of the Quranic verses series in answers to all questions received a score of 0.23 pAP, 0.32 F1, 0.19 EM, and 0.26 F1@1, while the text portion received 0.26 F1 and 0.29 EM.

### 7.4.4 Analysis and Discussion

This section analyses and discusses the results obtained from the answers generated by GPT-4 in response to the QUQA questions. Some incorrect answers appeared in the output, as illustrated in the following example. Only the part of the answer containing the error is displayed.

1. GPT-4 was unable to answer many questions, as illustrated in Example 4.

   **Example 4:**

   - **Question ID:** 328
   - **Question:**

بماذا اصطفى الله تعالى موسى عليه السلام على الناس في وقت زمانه ؟

What distinguished Allah, the Almighty, Moses (peace be upon him) over the people in his time?

- **Part of the GPT Answer:**

للاطلاع على جميع آيات القرآن الكريم التي تتعلق بكيفية اصطفاء الله تعالى موسى عليه السلام على الناس في وقت زمانه، يُفضل البحث عن التفاسير والشروح القرآنية المعتمدة. كوني نموذج لغة ذكاء اصطناعي، فإنني لا أملك القدرة على استخراج جميع الآيات القرآنية بدقة وفقط أستطيع تقديم إجابات عامة ومعرفية على أساس المعرفة المحدودة التي تم تدريبي عليها حتى عام ٢٠٢١.

To view all the verses of the Holy Qur'an that relate to how God Almighty chose Moses, peace be upon him, over other people during his time, it is preferable to search for approved Qur'anic interpretations and explanations. As an AI language model, I do not have the ability to accurately extract all the Quranic verses and can only provide general and epistemological answers based on the limited knowledge I have been trained on until 2021.

2. Example 5 shows the generation of incomprehensible words.

**Example 5:**

- **Question ID:** 485
- **Question:**

ما تأويل يوسف عليه السلام لرؤيا الملك ؟

What is the interpretation of the king's dream by Joseph, peace be upon him?

- **Part of the GPT Answer:**

. . . وهذا هو جواب يوسف (عليه السلام) على تفسير رؤيا الملك حسب الآيات (٤٣-٤٩) في سورة يوسف: قال يُوسُفُ **أَيُونَ التَّفْتَعِيْنَ سَبْعُون**

**بَقَرقَسُسُومَعَجَفِيْشأَصْدٌعيْنٌ الْقَدِيْأُوِنُ يَخَرِىٌّ وَالِّيَكَةُ ...**

. . .The answer of Yusuf (peace be upon him) to the interpretation of the king's dream according to the verses (43-49) in Surat Yusuf: Yusuf said **ayn altftan sbaon ygrgsomagfnsAsdanh.** ). . .

3. In Example 6, there is a contradiction, as it provides a specific answer and then negates it within the same text.

**Example 6:**

- **Question ID:** 326
- **Question:**

آنية الذهب محرم استخدامها على المسلم في الحياة الدنيا فهل هي كذلك في الآخرة ؟

Gold utensils are forbidden for use by Muslims in this life. Is the same true in the hereafter?

- **Part of the GPT Answer:**

**لا يوجد نص قرآني صريح يتحدث عن آنية الذهب في الآخرة.** ولكن يمكن استنتاج أن الملذات والزينة المحرمة في الحياة الدنيا قد تكون مباحة في الآخرة لأنها مكافأة للمؤمنين على عملهم الصالح. وفي القرآن الكريم هناك وصف للجنة وما فيها من نعيم. في سورة الزخرف (٤٣:٧١)، يقول الله تعالى: يُطَافُ عَلَيْهِم بِصِحَافٍ مِنْ ذَهَبٍ وَأَكْوَابٍ وَفِيهَا مَا تَشْتَهِيهِ الْأَنفُسُ وَتَلَذُّ الْأَعْيُنُ وَأَنتُمْ فِيهَا خَالِدُونَ . **هذه الآية تشير إلى أن اواني الذهبية ستستخدم لخدمة سكان الجنة، وستكون في قمة الرفاهية والترف.** . . .

**There is no explicit Qur'anic text that talks about vessels of gold in the hereafter**. However, it can be concluded that the pleasures and adornments that are forbidden in this life may be permissible in the hereafter because they are a reward for the believers for their good deeds. The Holy Qur'an describes heaven and its bliss.In

Surat Al-Zukhruf (43:71), God Almighty says: "To them will be passed round, dishes and goblets of gold: there will be there all that the souls could desire, all that their eyes could delight in: and ye shall abide therein (for eye)". **this verse indicates that golden utensils will be used to serve the inhabitants of Paradise**, and they will be at the height of luxury....

4. Example 7 shows an answer in which the text of the verse was correct, but the verse number and the name of the Surah were incorrect. The example indicates that this verse is from Surah Taha, verse 94, while in fact, it is from Surah Al-A'raf, verse 150. There were many answers in which the verse information was correct, but the text itself was incorrect.

**Example 7:**

- **Question ID:** 382
- **Question:**

لماذا أخذ موسى عليه السلام برأس أخيه هارون عليه السلام يجره إليه ؟

Why did Moses, peace be upon him, take the head of his brother Harun, peace be upon him, and drag him to Moses?

- **Part of the GPT Answer:**

... ﴿وَلَمَّا رَجَعَ مُوسَىٰ إِلَىٰ قَوْمِهِ غَضْبَانَ أَسِفًا قَالَ بِئْسَمَا خَلَفْتُمُونِي مِن بَعْدِي أَعَجِلْتُمْ أَمْرَ رَبِّكُمْ وَأَلْقَى الْأَلْوَاحَ وَأَخَذَ بِرَأْسِ أَخِيهِ يَجُرُّهُ إِلَيْهِ قَالَ ابْنَ أُمَّ إِنَّ الْقَوْمَ اسْتَضْعَفُونِي وَكَادُوا يَقْتُلُونَنِي فَلَا تُشْمِتْ بِيَ الْأَعْدَاءَ وَلَا تَجْعَلْنِي مَعَ الْقَوْمِ الظَّالِمِينَ﴾ **(طه: ٩٤)**

..."And when Moses returned to his people, angry and grieved, he said, "How wretched is that by which you have replaced me after [my departure]. Were you impatient over the matter of your Lord?" And he threw down the tablets and seized his brother by [the hair of] his head, pulling him toward him. [Aaron] said, "O son of my mother, indeed the people oppressed me and were about to kill me, so let not the enemies rejoice over me and do not place me among the wrongdoing people." **[Taha:94]**..

5. If GPT-4 found a matching word between the question and a verse, it retrieved the verse even when its meaning did not answer the question, as shown in Example 8.

**Example 8:**

- **Question ID:** 385
- **Question:**

لماذا تخلف **الأعراب** عن الخروج مع رسول الله صلى الله عليه وسلم إلى مكة ؟

Why did the **desert-Arabs** not go out with the Messenger of God, may God's prayers and peace be upon him, to Mecca?

- **Part of the GPT Answer:**

. . . وَمِنَ **الْأَعْرَابِ** مَن يُؤْمِنُ بِاللَّهِ وَالْيَوْمِ الْآخِرِ وَيَتَّخِذُ مَا يُنفِقُ قُرُبَاتٍ عِندَ اللَّهِ وَصَلَوَاتِ الرَّسُولِ أَلَا إِنَّهَا قُرْبَةٌ لَّهُمْ سَيُدْخِلُهُمُ اللَّهُ فِي رَحْمَتِهِ إِنَّ اللَّهَ غَفُورٌ رَّحِيمٌ . . .

. . . "But some of the **desert Arabs** believe in Allah and the Last Day, and look on their payments as pious gifts bringing them nearer to Allah and obtaining the prayers of the Messenger. Aye, indeed they bring them nearer (to Him): soon will Allah admit them to His Mercy: for Allah is Oft-forgiving, Most Merciful"...[ at-Tawbah:99]. . .

- **Golden Answer:**

قل فمن يملك لكم من الله شيئا إن أراد بكم ضرا أو أراد بكم نفعا بل كان الله بما تعملون خبيرا ١١ بل ظننتم أن لن ينقلب الرسول والمؤمنون إلى أهليهم أبدا وزين ذلك في قلوبكم وظننتم ظن السوء وكنتم قوما بورا ١٢ الفتح .

"Then who could prevent Allah at all if He intended for you harm or intended for you benefit? Rather, ever is Allah, with what you do, Acquainted [11]. But you thought that the Messenger and the believers would never return to their families, ever, and that was made pleasing in your hearts. And you assumed an assumption of evil and became a people ruined."[12].[Sura Al-Fat'h].

6. In Example 9, it failed to answer the question when the words in the question did not match the words in the correct answer verses.

   **Example 9:**

   - **Question ID:** 534
   - **Question:**

     ما علامة **إتيان** العذاب لقوم نوح عليه السلام؟
     What is the sign of the **torment coming** to the people of Noah, peace be upon him?

   - **Part of the GPT Answer:**

     . . . في سورة هود ... في الآية ٤٤ وَقِيلَ يَا أَرْضُ ابْلَعِي مَاءَكِ وَيَا سَمَاءُ أَقْلِعِي وَغِيضَ الْمَاءُ وَقُضِيَ الأَمْرُ وَاسْتَوَتْ عَلَى الْجُودِيِّ وَقِيلَ بُعْدًا لِلقَوْمِ الظَّالِمِينَ . . .

     ...And it was said, "O Earth, swallow your water, and O sky, withhold [your rain]." And the water subsided, and the matter was accomplished, and the ship came to rest on the [mountain of] Judiyy. And it was said, "Away with the wrongdoing people."[Hud:44]...

   - **Golden Answer:**

     العلامة هي :نبع الماء بقوة من التنور ـ وهو المكان الذي يخبز فيه علامة على مجيء العذاب. **حَتَّى إِذَا جَاء أَمْرُنَا** وَفَارَ التَّنُّورُ قُلْنَا احْمِلْ فِيهَا مِن كُلٍّ زَوْجَيْنِ اثْنَيْنِ وَأَهْلَكَ إِلاَّ مَن سَبَقَ عَلَيْهِ الْقَوْلُ وَمَنْ آمَنَ وَمَا آمَنَ مَعَهُ إِلاَّ قَلِيلٌ ٤٠

**هود.**

The sign is: water gushing powerfully from the oven - which is the reported location - is a sign of the coming of torment. "At length, behold! there came **Our command**, and the fountains of the earth gushed forth! We said: "Embark therein, of each kind two, male and female, and your family - except those against whom the word has already gone forth,- and the Believers." but only a few believed with him."

7. It also faced difficulties when a word in the question and a verse were similar, but differed in meaning due to context, as shown in Example 10. This led to incorrect processing of the question. For example, the word " **شعب** " in Arabic means both “people” and “branches.” In Example 10, the question refers to “branches of faith” based on the context of the sentence, but GPT-4 processed it as equivalent to “faithful people.

**Example 10:**

- **Question ID:** 1892
- **Question:**

اذكر بعض شعب الإيمان؟

Describe some **branches of faith**?

- **Part of the GPT Answer:**

. . . في القرآن الكريم تم الإشارة إلى العديد من **شعوب الإيمان** منهم: ١.
بنو إسرائيل: . . .

...In the Holy Qur’an, **many peoples of faith**are referred to, including: 1. The Children of Israel...

- **Golden Answer:**

**شعب الايمان هي** الإيمان بالله، وملائكته، وكتبه، ورسله، واليوم الآخر ...

... **Branches of faith** include belief in God, His angels, His books, His messengers, and the Last Day...

8. In Example 11, the answer was composed of a series of verses. This suggests that the model found it difficult to provide a specific answer.

   **Example 11:**

   - **Question ID:** 291
   - **Question:**

     اذكر نسبة دخول الخلق للجنة من مرتبة السابقون؟

     Describe the **percentage** of people entering heaven from the **forerunners** rank?

   - **Part of the GPT Answer:**

     لا يوجد **رقم محدد** لنسبة دخول الخلق للجنة من مرتبة **السابقون** في القرآن الكريم ...

     There is **no specific number** for the percentage of people entering Paradise from the rank of the "**forerunners** "in the Holy Quran....

   - **Golden Answer:**

     ... **وَالسَّابِقُونَ السَّابِقُونَ** ١٠ أُوْلَئِكَ الْمُقَرَّبُونَ ١١ فِي جَنَّاتِ النَّعِيمِ ١٢ **ثُلَّةٌ مِّنَ الْأَوَّلِينَ** ١٣ وَقَلِيلٌ مِّنَ الْآخِرِينَ ١٤ الواقعة.

     ..."And **the forerunners, the forerunners** [10]. These are the nearest [11]. In the Gardens of Bliss [12]. **A group from the ancients** [13]. And **a few from the latecomers** [14]."[Surah Al-Waqia]

In addition to the above examples, many answers did not mention Quranic verses, such as in Question 1. Other GPT-4 answers contain many related topics but do not directly answer the question, such as in Question 487.

## 7.5 Experiment 2: Answering Questions from the Hadith

### 7.5.1 Datasets

This experiment used the first 601 questions from HAQA. The types of questions in this corpus are confirmation (97 questions), descriptive (387 questions), and factual (117 questions).

### 7.5.2 Evaluation

The output generated by GPT-4 in response to the HAQA questions contains natural-sounding text and one or more hadiths, as shown in Example 12.

**Example 12:**

- **Question ID:** 598
- **Question:**

  متى كان النبي يعتكف ؟

  When did the Prophet, may God bless him and grant him peace, seclude himself?

- **The GPT Answer:**

  النبي صلى الله عليه وسلم كان يعتكف في العشر الأواخر من شهر رمضان، وفي بعض الأحيان يمتد الاعتكاف إلى أيام أخرى. وفي حديث عائشة رضي الله عنها قالت: كان النبي صلى الله عليه وسلم يعتكف العشر الأواخر من رمضان حتى توفاه الله. (البخاري)

  The Prophet, peace and blessings of God be upon him, used to seclude himself during the last ten days of the month of Ramadan, and in some cultures, he secluded himself during other days. The wife of the Prophet) The Prophet, peace and blessings of God be upon him, used to practice I'tikaf in the last ten days of Ramadan till he died. (Bukhari)

The answers to the Hadith were evaluated similarly to those of the Holy Quran because I noted the same observations. For example, I manually evaluated the text portion because, at times, the GPT-4 answer is correct but uses synonyms for the answers in the gold dataset, as shown in Example 13.

**Example 13:**

- **Question ID:** 481
- **Question:**

  ماذا يفعل المصلي إذا أراد أحد أن يمر بينه وبين السترة ؟

  What should a worshipper do if someone wants to pass between him and the sutrah?

- **Part of the GPT Answer:**

  يجب على المصلي منع الشخص الذي يريد المرور بينه وبين السترة أثناء المصلي يؤدي صلاته

  The worshipper must prevent anyone from passing between him and the sutrah while he is performing his prayer.

- **Gold Answer:**

  يرده .

  Return him.

The Hadith portion of the answer was evaluated using https://sunnah.com/ as well as the Leeds University and King Saud University (KSU) Hadith Corpus (Altammami *et al.*, 2020).

### 7.5.3 Results

This section presents the performance of GPT-4 with the HAQA dataset. Three experiments were performed for the HAQA dataset, similar to QUQA. In each experiment, I entered a particular type of question. The HAQA contains three kinds of questions: confirmation (259 questions), descriptive (862 questions), and fact (245 questions). Table 7.3 shows a summary of the results for the Hadith series portion of the GPT-4 answers, while Table 7.4 shows the results for the produced-text portion.

In this study, I used the same four evaluation metrics—EM, pAP, F1, and F1@1 scores—as in Experiment 1 of this chapter. When evaluating the natural text and Hadith portions of the answers, I treated them as bags of words, similar to the method used in Experiment 1 and Chapter 6.

In general, the performance of GPT-4 in answering questions with evidence from the Hadith was worse than in finding evidence from the Quran, achieving 0.137 F1@1, 0.16 EM, 0.135 F1, and 0.140 pAP, compared to 0.26 F1@1, 0.19 EM, 0.32 F1, and 0.23 pAP, as shown in Table 7.3 and Table 7.1. The answers to confirmation questions from Hadith obtained the best results: 0.154 F1@1, 0.205 EM, 0.170 F1, and 0.170 pAP. The answers to descriptive questions were next, with results of 0.165 F1@1, 0.192 EM, 0.157 F1, and 0.167 pAP. Finally, the answers to fact-type questions achieved the following results: 0.079 F1@1, 0.093 EM, 0.079 F1, and 0.079 pAP.

For the entire regular text of the GPT-4 answer to the Hadith questions, GPT-4 achieves good results, with a 0.405 F1 score, as shown in Table 7.4. This result is approximately twice that of the produced text portion of the GPT-4 answer to questions related to the Quran, which achieved a 0.26 F1 score. The performance of GPT-4 when answering descriptive questions is the best, with a 0.431 F1 score, followed by the fact question type with a 0.391 F1 score. Finally, the confirmation question type achieved a 0.338 F1 score. The EM result of the produced text portion of the GPT-4 answers was disappointing, achieving 0.008 EM for all types of questions.

### 7.5.4 Analysis and Discussion

This section analyses and discusses the results obtained from the answers generated by GPT-4 in response to the HAQA questions. As illustrated in the following examples, some incorrect answers appeared in the output.

| Question Type | F1@1 | EM | F1 | pAP |
|---|---|---|---|---|
| **All** | 0.137 | 0.16 | 0.135 | 0.140 |
| **Confirmation** | 0.154 | **0.205** | **0.170** | **0.170** |
| **Descriptive** | **0.165** | 0.192 | 0.157 | 0.167 |
| **Fact** | 0.079 | 0.093 | 0.079 | 0.079 |

Table 7.3: The evaluation results of the Hadith portion of the GPT-4 answers.

| Question Type | EM | F1 |
|---|---|---|
| **All** | 0.008 | 0.405 |
| **Confirmation** | 0.012 | 0.338 |
| **Descriptive** | 0.013 | **0.431** |
| **Fact** | **0.040** | 0.391 |

Table 7.4: The evaluation results of the produced-text portion of the GPT-4 answers.

1. GPT-4 answers the question correctly, but it fails to find evidence from the Hadith, as shown in Example 14.

   **Example 14:**

   - **Question ID:** 392
   - **Question:**

     كم كان عمر عمر بن الخطاب رضي الله عنه حين قتل ؟

     How old was Omar bin Al-Khattab, may God be pleased with him, when he was killed?

   - **Part of the GPT Answer:**

     عمر بن الخطاب رضي الله عنه كان عمره ٦٣ عامًا عندما قتل في الهجرة النبوية الشريفة. لا يوجد حديث نبوي يذكر عمر الخليفة عمر بن الخطاب رضي الله عنه عند وفاته بشكل مباشر. إنما يمكن حساب عمره بناءً على معلومات أخرى موثوقة من التاريخ. عمر بن الخطاب رضي الله عنه ولد في عام ٥٨٣ ميلادي وقتل في الحادي والعشرين من ذو الحجة عام ٢٣ هـ ٢)

نوفمبر ٦٤٤ ميلادي).

Omar bin Al-Khattab, may God be pleased with him, was 63 years old when he was killed during the noble Prophet's migration. There is no Prophetic hadith that directly mentions the age of Caliph Omar bin Al-Khattab, may God be pleased with him, at the time of his death. However, his age can be calculated based on other reliable historical information. Omar bin Al-Khattab, may God be pleased with him, was born in the year 583 AD and was killed on the twenty-first of Dhul-Hijjah in the year 23 AH (November 2, 644 AD).

- **Gold Answer:**

ثلاث وستون سنة والدليل حديث أبي إسحاق قال كنت جالسا مع عبد الله بن عتبة فذكروا سني رسول الله فقال بعض القوم كان أبو بكر أكبر من رسول الله قال عبد الله قبض رسول الله وهو ابن ثلاث وستين ومات أبو بكر وهو ابن ثلاث وستين وقتل عمر وهو ابن ثلاث وستين قال فقال رجل من القوم يقال له عامر بن سعد حدثنا جرير قال كنا قعودا عند معاوية فذكروا سني رسول الله فقال معاوية قبض رسول الله وهو ابن ثلاث وستين سنة ومات أبو بكر وهو ابن ثلاث وستين وقتل عمر وهو ابن ثلاث وستين رواه مسلم

Sixty-three years, and the evidence of Abu Ishaq said: I was with Abdullah bin Utbah, so they mentioned the years of the Messenger of God, peace and blessings of God be upon him, in a hadith, and some of the people said: Abu Bakr was older than the Messenger of God, peace and blessings be upon him. Abdullah said: Allah's Messenger (Peace be upon him) died when he was sixty-three, and Abu Bakr died when he was sixty-three and so 'Umar fell as a martyr when he was sixty-three. A person from the people who was called 'Amir bin Sa'd reported that Jabir had said: We were sitting with Mu'awiya that there was a discussion about the age of Allah's Messenger (Peace be upon him). Thereupon Mu'awiya said: Allah's Messenger (Peace be upon him) died when he had attained the age of sixty-three, and Abu Bakr died when he had attained the age of sixty-three, and Umar fell as a martyr when he had attained the age of sixty-three." Narrated by Muslim.

2. Example 15 shows that GPT retrieves the Hadith correctly but switches between two words, completely changing the meaning.

   **Example 15:**

   - **Question ID:** 536
   - **Question:**

     ما هي أفضل صفوف النساء ؟

     What are the best rows for women?
   - **Part of the GPT Answer:**

     يقول النبي محمد صلى الله عليه وسلم في الحديث الشريف: **خير صفوف النساء أولها** وشرها آخرها. (رواه مسلم) هذا الحديث يشير إلى أن أفضل صفوف النساء في الصلاة هو الصف **الأول** وأقلهم شرفًا هو الأخير.

     The Prophet Muhammad, may God bless him and grant him peace, says in the noble hadith: "**The best of the woman's row is the first** and the worst of their rows is the last.." (Narrated by Muslim) This hadith indicates that the best row for women in prayer is the first row, and the least honorable is the last.
   - **Gold Answer:**

     آخرها والدليل حديث أبي هريرة رضي الله عنه قال قال رسول الله صلى الله عليه وسلم خير صفوف الرجال أولها وشرها آخرها **وخير صفوف النساء آخرها** وشرها أولها رواه مسلم

     The last of them, and the evidence, is the hadith of Abu Hurairah, may God be pleased with him, who said that the Messenger of God, may God's prayers and peace be upon him, said: "The best of the men's rows (in Salat) is the first row and the worst is the last; **but the best of the woman's row is the last** and the worst of their rows is the first." Narrated by Muslim.

3. In Example 16, GPT changed one word in the Hadith, completely affecting its meaning. This word is the keyword in the Hadith.

   **Example 16:**

   - **Question ID:** 685
   - **Question:**

     من هم أطول الناس أعناقا يوم القيامة ؟

     Who will have the longest necks on the Day of Resurrection?

   - **Part of the GPT Answer:**

     في الحديث النبوي الشريف، روى أن رسول الله صلى الله عليه وسلم قال: ءطول الناس أعناقاً يوم القيامة **المنافقون** آ رواه مسلم في صحيحه. هذا الحديث يشير إلى أن المنافقين سيكونون أطول الناس أعناقاً يوم القيامة، وذلك استدلالاً من الحديث الشريف. وهذا التفسير يعني أن المنافقين سيكونون أكثر وضوحاً وظلماً وهم من يظهرون الإيمان بالإسلام ولكن قلوبهم تكابر وتعصي الله ورسوله. ويوم القيامة سيجازون على نفاقهم وتضليلهم.

     In the noble Prophetic hadith, it was narrated that the Messenger of God, may God bless him and grant him peace, said: "The people with the longest necks on the Day of Resurrection are **the hypocrites**." Narrated by Muslim in his Sahih. This hadith indicates that the hypocrites will have the longest necks on the Day of Resurrection, as inferred from the noble hadith. This interpretation suggests that the hypocrites will be more visible and unjust, as they are those who outwardly show belief in Islam while their hearts are arrogant and they disobey God and His Messenger. On the Day of Resurrection, they will be held accountable for their hypocrisy and misguidance.

- **Gold Answer:**

المؤذنون والدليل حديث معاوية رضي الله عنه قال سمعت رسول الله يقول **المؤذنون** أطول الناس أعناقا يوم القيامة رواه مسلم

The mu'adhdhins and the evidence is Mu'awiya said that he heard God's Messenger say, "**The mu'adhdhins** will have the longest necks on the day of resurrection." Narrated by Muslim.

4. Sometimes, it is recommended to review books on the Islamic religion to find evidence, as shown in Example 17.

**Example 17:**

- **Question ID:** 607
- **Question:**

اذكر مواقيت الحج المكانية ؟

Mention the location of Miqats of the Hajj ?

- **Part of the GPT Answer:**

مواقيت الحج المكانية تعتبر الحدود الجغرافية التي يجب على المسلم تجاوزها بالإحرام قبل قصد الحج أو العمرة. وتسمى هذه المواقيت بـ الميقاتْ. والميقات المذكورة في الأحاديث النبوية الشريفة هي خمسة: الأول: ذو الحليفة: وهو الميقات لأهل المدينة ومن ستقبل طريقهم ويقع على بعد حوالي ٤٠٠ كم شمال مكة. الثاني: الجحفة: وهو الميقات لأهل مصر والمغرب والسودان والاندلس ومن ستقبل طريقهم من المسلمين وهي تقع الآن في ميسعر. الثالث: قرن المنازل: وهو الميقات لأهل النجد. الرابع: يلملم: وهو الميقات لأهل اليمن وهو يبعد أكثر من ٢٠٠ كيلومترا إلى الجنوب الشرقي من مكة المكرمة. الخامس: ذات عرق: وهو الميقات لأهل العراق ويقع في الغرب

الجنوبي للعراق. ويُذكر أن هذا السؤال يتعلق بالفقه الإسلامي ومعرفة الأحاديث النبوية الشريفة عن هذا الموضوع؛ لذا ننصح بمراجعة كتب الفقه والتفسير والحديث للحصول على معرفة أكثر تفصيلية حول الموضوع.

The spatial Miqat boundaries are considered the geographical boundaries that a Muslim must cross while entering into ihram before intending to perform Hajj or Umrah. These boundaries are referred to as "Miqat." The Miqats mentioned in the noble Prophetic hadiths are five: the first is Dhu al-Hulayfa, which is the Miqat for the people of Medina and those traveling to Mecca, located about 400 km north of the city. The second is Al-Juhfa, the Miqat for the people of Egypt, Morocco, Sudan, Andalusia, and other Muslims traveling to the area; it is now located in Maysa'ar. The third is Al-Qarn Al-Manazel, which is the Miqat for the people of Al-Najd. The fourth is Yalamlam, the Miqat for the people of Yemen, situated more than 200 kilometers to the southeast of Mecca. The fifth is Dhat Irq, which serves as the Miqat for the people of Iraq and is located in the southwestern part of Iraq. It is noteworthy that this topic relates to Islamic jurisprudence and knowledge of the noble Prophetic hadiths on this subject. Therefore, we do not recommend reviewing books on jurisprudence, interpretation, and hadith to obtain more detailed knowledge about the topic.

- **Gold Answer:**

هي لأهل المدينة ذو الحليفة لأهل الشام الجحفة لأهل نجد قرن المنازل لأهل اليمن يلملم لأهل العراق ذات عرق. والدليل حديث ابن عباس رضي الله عنه قال إن النبي  وقت لأهل المدينة ذا الحليفة ولأهل الشأم الجحفة ولأهل نجد قرن المنازل ولأهل اليمن يلملم رواه البخاري ومسلم

It is for the people of Medina: Dhu Al-Hulayfa; for the people of the Levant: Al-Juhfa; for the people of Najd: Qarn Al-Manazil; for the people of Yemen: Yalamlam; and for the people of Iraq: Dhat Irq. The evidence is the hadith of Ibn Abbas, may God be pleased with him, who said that the Prophet, peace and blessings of God be upon him, had fixed Dhul Hulaifa as the Miqat for the people of Medina; Al-Juhfa for the people of Sham; and Qarn Ul-Manazil for the people of Najd; and Yalamlam for the people of Yemen Narrated by al-Bukhari and Muslim.

5. In Example 18, GPT understood the term النشرة as spreading الانتشار , but its intended meaning in Islamic context refers to removing magic from the bewitched, حل السحر عن المسحور .

   **Example 18:**

   - **Question ID:** 92
   - **Question:**

   ما حكم النشرة ؟

   What is the ruling on the **nushra**?

   - **Part of the GPT Answer:**

   الحديث النبوي الشريف يحث على البحث عن العلم والمعرفة **ونشرها** بين الناس، ويعتبر حديث آبلغوا عني ولو آيةْ من الأحاديث التي تشجع على نشر العلم. قال رسول الله صلى الله عليه وسلم: آبلغوا عني ولو آية، وحدثوا عن بني إسرائيل ولا حرج، ومن كذب علي متعمدًا فليتبوأ مقعده من النارْ (رواه البخاري). وفي حديث آخر، قال رسول الله صلى الله عليه وسلم: آمن دل على خير فله مثل أجر فاعلهْ (رواه مسلم). تشير هذه الأحاديث إلى حث النبي صلى الله عليه وسلم على نشر العلم والخير بين الناس، وتعليم الناس ما ينفعهم في دينهم. ولكن يجب أن يؤخذ العلم من مصادره الصحيحة والموثوقة وتفقيه الناس في الدين على أصوله، وينبغي أن يكون النشر حكيماً وموجهاً للخير والصلاح، وتجنب نشر الأخبار الكاذبة أو ما يزعزع استقرار المجتمع وينشر الفتنة. في المجمل، يعتبر نشر العلم والمعرفة وتوجيه الناس لما يحبه الله

ويرضاه أمراً مستحباً وفرض حسب المقدرة والظروف.

The noble Prophetic hadith urges the search for knowledge and the **spreading** of it among people. The hadith 'Convey even a verse on my authority' is one of the hadiths that encourages the spread of knowledge. The Messenger of God, may God bless him and grant him peace, said: "Pass on information from me, even if it is only a verse of the Qur'an; and relate traditions from the Bani Isra'il, for there is no restriction; but let him who deliberately lies against me come to his abode in hell." (Narrated by Al-Bukhari). In another hadith, the Messenger of God, may God bless him and grant him peace, said: "He who guides (others) to an act of goodness, will have a reward similar to that of its doer." (Narrated by Muslim). These hadiths emphasize the Prophet's encouragement, may God bless him and grant him peace, to spread knowledge and goodness among people and to teach them what benefits them in their religion. However, knowledge must be drawn from correct and reliable sources, and people must be educated in religion according to its principles. The dissemination of knowledge should be wise and directed toward goodness and righteousness. Why publish false information or anything that destabilizes society and spreads strife? In general, spreading knowledge and understanding about what God loves and is pleased with is considered desirable and obligatory according to the situation and circumstances.

- **Gold Answer:**

إذا كانت بسحر فلا يجوز والدليل عن جابر بن عبد الله رضي الله عنه قال سئل رسول الله عن **النشرة** فقال هو من عمل الشيطان رواه أحمد وأبو داود وصححه الألباني كما في المشكاة

If it was due to witchcraft, then it is not permissible. The evidence is on the authority of Jabir told that when the Prophet was asked about a charm for one who is possessed **(nushra)*** he replied, "It pertains to the work of the devil". Narrated by Ahmad and Abu Dawud, and authenticated by Al-Albani, as in Al-Mishkat.

6. GPT mentioned a large number of fake hadiths while answering questions, which reflects its weakness on this point, as shown in Example 19.

   **Example 19:**

   - **Question ID:** 324
   - **Question:**

     من هو الذي أوتي مزمارا من مزامير آل داود ؟

     Who among the Mazamir of Prophet Dawud (David) was given a Mizmar (sweet melodious voice)?
   - **Part of the GPT Answer:**

     الجواب: في الحديث النبوي الشريف، يقال أن الملك المسيح الذي سيرسله الله إلى الأرض، والذي سيهدي الحق ويكون من المؤمنين، هو الذي سيمنحه الله بمزمار من مزامير آل داود. (رواه الإمام أحمد في مسنده والألباني في السلسلة الصحيحة)

     In the noble Prophetic hadith, it is clear that the Jesus King, whom God will send to earth to guide the truth and the believers, is the one whom God will bestow with a Mizmar (sweet melodious voice) from the Mazamir of Prophet Dawud (David). (Narrated by Imam Ahmad in his Musnad and by Al-Albani in Sharia law).
   - **Gold Answer:**

     أبو موسى الأشعري رضي الله عنه والدليل عن أبي موسى الأشعري رضي الله عنه أن رسول الله صلى الله عليه وسلم قال له : لقد أتيت مزمارا من مزامير آل داود . رواه البخاري ومسلم

     Abu Musa Al-Ash'ari. The evidence is Abu Musa Al-Ash'ari (May Allah be pleased with him) reported: The Messenger of Allah (Peace be upon him) said to him, "You have been given a Mizmar (sweet melodious voice) out of the Mazamir of Prophet Dawud (David)."

7. The answers sometimes include words from languages other than Arabic, as shown in Example 20.

**Example 20:**

- **Question ID:** 651
- **Question:**

ما هي البركة ؟

What is blessing?

- **Part of the GPT Answer:**

البركة هي الزيادة والخير الذي يمنحه الله لشخص أو موضوع أو مكان. عندما تكون هناك بركة في شيء، فإن الله يُقدِّم له خير ونماء ويجعله مُتعدّدًا ومفيدًا. يشير الحديث النبوي الشريف إلى العديد من التوجيهات التي تساعد المسلمين في تحقيق البركة في حياتهم وأعمالهم. في الحديث النبوي الشريف، روى الإمام البخاري عن عمرو بن عوف رضي الله عنه قال: قال رسول الله صلى الله عليه وسلم : ألجمعة بالجمعة ، والرمضان إلى الرمضان ، والعمرة إلى العمرة كفارة لما بينهن إذا اجتنب الكبائر، والصلوات الخمس ، والجمعة إلى الجمعة ، والجمعة إلى الجمعة مكفرات للوب إذا اجتنب الكبائر
**Послание от Господа**
مُتَعدّدَ ومفـدْ (صحيح البخاري: ٢٦٢٤) في هذا الحديث، يُشير رسول الله صلى الله عليه وسلم إلى الجمعة وصوم رمضان والعمرة على أنها وسائل لكفّارة الذنوب وتحقيق البركة.

Blessing is the increase and goodness that God gives to a person, subject, or place. When there is blessing in something, God brings goodness and growth to it, making it abundant and useful. The Noble Prophet's Hadith indicates many ways that help Muslims achieve blessings in their lives and deeds. In a noble Prophetic hadith, Imam Al-Bukhari narrated on the authority of Amr bin Awf, may God be pleased with him, who said: The Messenger of God, may God bless him and grant him peace, said: 'Friday for Friday, and Ramadan to Ramadan, and Umrah to Umrah is expiation for what comes between them if one avoids major sins, and the five daily prayers, and Friday to Friday, and Friday after Friday is an expiation for the sinner if he avoids major sins. **Послание от Господа** mnaDD w mfD(Sahih al-Bukhari: 2624) In this hadith, the Messenger of God, may God bless him and grant him peace, refers to Friday, Ramadan fasting, and Umrah as means of expiating sins and attaining blessings.

- **Gold Answer:**

ثبوت الخير الإلهي في الشيء. والدليل حديث عَنْ جَابِرِ بْنِ عَبْدِ اللَّهِ ـ رضى الله عنهما ـ هَذَا الْحَدِيثَ قَالَ قَدْ رَأَيْتُنِي مَعَ النَّبِيِّ صلى الله عليه وسلم وَقَدْ حَضَرَتِ الْعَصْرُ وَلَيْسَ مَعَنَا مَاءٌ غَيْرَ فَضْلَةٍ فَجُعِلَ فِي إِنَاءٍ، فَأُتِيَ النَّبِيُّ صلى الله عليه وسلم بِهِ فَأَدْخَلَ يَدَهُ فِيهِ وَفَرَّجَ أَصَابِعَهُ ثُمَّ قَالَ حَىَّ عَلَى أَهْلِ الْوُضُوءِ، الْبَرَكَةُ مِنَ اللَّهِ . فَلَقَدْ رَأَيْتُ الْمَاءَ يَتَفَجَّرُ مِنْ بَيْنِ أَصَابِعِهِ، فَتَوَضَّأَ النَّاسُ وَشَرِبُوا، فَجَعَلْتُ لاَ آلُو مَا جَعَلْتُ فِي بَطْنِي مِنْهُ، فَعَلِمْتُ أَنَّهُ بَرَكَةٌ. قُلْتُ لِجَابِرٍ كَمْ كُنْتُمْ يَوْمَئِذٍ قَالَ أَلْفًا وَأَرْبَعَمِائَةٍ. تَابَعَهُ عَمْرٌو عَنْ جَابِرٍ. وَقَالَ حُصَيْنٌ وَعَمْرُو بْنُ مُرَّةَ عَنْ سَالِمٍ عَنْ جَابِرٍ خَمْسَ عَشْرَةَ مِائَةً. وَتَابَعَهُ سَعِيدُ بْنُ الْمُسَيَّبِ عَنْ جَابِرٍ.

Proof of divine goodness in a thing. The evidence is the hadith of Jabir bin Abdullah, may God be pleased with him. He said, " I was with the Prophet (Peace be upon him) and the time for the 'Asr prayer became due. We had no water with us except a little which was put in a vessel and was brought to the Prophet (Peace be upon him). He put his hand into it and spread out his fingers and then said, "Come along! Hurry up! All those who want to perform ablution. The blessing is from Allah." I saw the water gushing out from his fingers. So the people performed the ablution and drank, and I tried to drink more of that water (beyond my thirst and capacity), for I knew that it was

> a blessing. The sub-narrator said: I asked Jabir, "How many persons were you then?" He replied, "We were one-thousand four hundred men." Narrated by Al-Bukhari.

This chapter answers the fourth research question, **RQ4:** "Is GPT-4 a Good Islamic Expert for Answering Quran Questions?" Based on the results, problems, and limitations identified, GPT-4 requires further development and research to improve the quality of their performance.

## 7.6 Conclusion

This chapter reviewed research examining LLMs' ability to answer questions. The results indicated good performance in some domains and languages, such as English. Next, the performance of the GPT-4 model in answering questions related to the Holy Quran and the Hadith in CA was investigated. The experiment revealed generally weak performance from GPT-4. The performance of GPT-4 with Quran questions was better than that with Hadith questions in the evidence portion of the answers. In the produced-text portion of the answers, the performance of GPT-4 with Hadith questions was better than that with Quran questions. The findings of this study highlight the model's limitations regarding the CA language and emphasize the need for further research in this area and more training data.

# Chapter 8

# Improving the Large Language Models (LLMs) (GPT-4) Models using Retrieval-Augmented Generation Technique

## 8.1 Introduction

One of the challenges in NLP concerns answering questions, and studies have tended to use LLMs for this task. However, these models often generate misleading and incorrect answers. Answering religious questions is a delicate and sensitive topic. Therefore, this chapter contributes to improving the performance of LLMs in Arabic by using the RAG technique. It compares the answers of standard GPT-4 with those of GPT-4 using the RAG technique for the same questions. This chapter was published in (Alnefaie *et al.*, 2024).

## 8.2 Background

LLMs have shown great promise in generation tasks and in understanding natural language (Brown *et al.*, 2020; Chowdhery *et al.*, 2023; Ouyang *et al.*, 2022). Many studies have focused on investigating the performance of LLMs, especially GPT, on various downstream tasks. However, GPT-4 faces challenges when answering Quranic questions in Arabic. Unfortunately, the performance of GPT-4 has been found to be poor in this respect (Alnefaie *et al.*, 2023b).

LLMs are trained on massive and specific datasets. Therefore, their answers to questions in a specific field, such as the Islamic religion, often suffer from issues such as memory efficiency of parametric knowledge (Heinzerling & Inui, 2020), outdated parametric knowledge (Liska *et al.*, 2022), and hallucinations (Bang *et al.*, 2023; Borji, 2023).

RAG can be used to address these limitations. RAG is a technique that takes advantage of the capabilities of LLMs to answer questions in a specific domain. This technology aims to provide an LLM with a knowledge base, from which the model retrieves answers to questions to improve its accuracy (Gao *et al.*, 2023).

The research question, as mentioned before, is: **RQ5** Does using the RAG technique improve the performance of GPT-4 in answering Quranic questions in Arabic? To answer this question, a set of Quranic questions was posed to both standard GPT-4 and GPT-4 using RAG, and the results were compared.

## 8.3 Related Work

Studies have aimed to evaluate GPT's performance in answering different types of questions. Several studies have shown good performance of GPT. For more details, please refer to the section 7.2.

Several studies have investigated the performance of GPT in answering Islamic questions. Sembok & Wani (2023) evaluated 19 Islamic questions in English, finding that GPT-3 answered only nine questions correctly. Alnefaie *et al.* (2023b) proposed evaluating GPT's performance in answering 2,189 Quranic questions in Arabic, with GPT-4 achieving the following results: 0.26 F1@1, 0.19 EM, 0.32 F1, and 0.23 pAP. Its details are explained in Chapter 7. Rizqullah *et al.* (2023) assessed 66 Islamic questions in Indonesian, with results showing that GPT-4 achieved 0.42 EM and 0.30 F1. We can conclude from these studies that GPT needs improvement in answering Islamic questions.

Several studies have examined the extent of the RAG technique's effect on improving GPT's performance in answering questions. Lála *et al.* (2023) proposed using RAG with GPT-4 to answer 55 scientific questions, utilizing a collection of scientific papers as the external knowledge base. They concluded that using this technique improved the results of the LLM. Zakka *et al.* (2024) recommended using RAG in the clinical decision-making field. They used GPT-4 with the RAG technique to answer 314 questions, noting a noticeable improvement in the results compared to using standard LLMs. Di Palma (2023) proposed using RAG with GPT-3.5 as a recommender system. Zhang *et al.* (2023) suggested using RAG for financial sentiment analysis, exploring the technique's performance with financial datasets from news and social media. In this case, RAG increased the LLM's F1 performance and accuracy by 48% and 15%, respectively. These studies unanimously agree that RAG enhances the performance of LLMs in various fields.

To the best of our knowledge, only one study has investigated the performance of RAG in the religious domain. Yusuf Alan *et al.* (2024) developed an Islamic question–answering model for Turkish. The external knowledge base included Turkish translations of the Quran, interpretations of the Quran, and Turkish Hadith. This study presented the RAG and standalone GPT-3.5 answers to only three questions, two of which were confirmation questions and one descriptive, but it did not present any results. To our knowledge, no study has thoroughly investigated the performance of RAG with Islamic questions using different question types and varying degrees of difficulty in Arabic.

## 8.4 Proposed Model

This study enhanced the LLM by using the RAG technique. The RAG architecture is shown in Fig. 8.1 (Ranjit *et al.*, 2023; Wang *et al.*, 2023a). It consists of two stages:

1. **Indexing Stage:**

   It was an offline stage with the following four main components:

   - **Loading:** The documents or knowledge sources needed to be loaded—in this case, the source of knowledge was Quranic text [1].

[1] http://tanzil.net/

- **Splitting:** The text was split into smaller chunks, which facilitated the search process since searching a small chunk is easier and faster than searching a large one. This step was also useful for indexing. In this experiment, the chunk size was set to 1000.
- **Embedding:** A vector representation was created for each chunk using the embedding model. Converting text into a vector simplified the semantic search process, as the search was for the vector most similar to it. The OpenAI embedding model was used [1].
- **Storing:** These vectors were stored in the FAISS [2] vector store.

2. **Retrieval and Generation Stage:**

   It was an online stage. First, the user's question was embedded using the same embedding model, and the most relevant chunks were retrieved from the index. Then, the LLM generated the answer to the question using the relevant data.

The LangChain framework [3] was used to implement the RAG, and Google Colab [4] was used to run the experiment. GPT-4 was the LLM used in this experiment. The code is publicly available [5].

## 8.5 Experiment: Answering Questions from the Holy Quran

### 8.5.1 Dataset

This study used the QUQA dataset, with the answers consisting of two parts: the first being scholars' answers, and the second being verses from the Holy Quran that contained the answers. I used 804 questions from this dataset to conduct

[1]https://python.langchain.com/docs/concepts/embedding$_{m}odels/$
[2]https://python.langchain.com/docs/integrations/vectorstores/faiss/
[3]https://python.langchain.com/docs/introduction/
[4]https://colab.research.google.com/
[5]https://github.com/scsaln/RAG/blob/main/QuranlangchainGPT4.ipynb

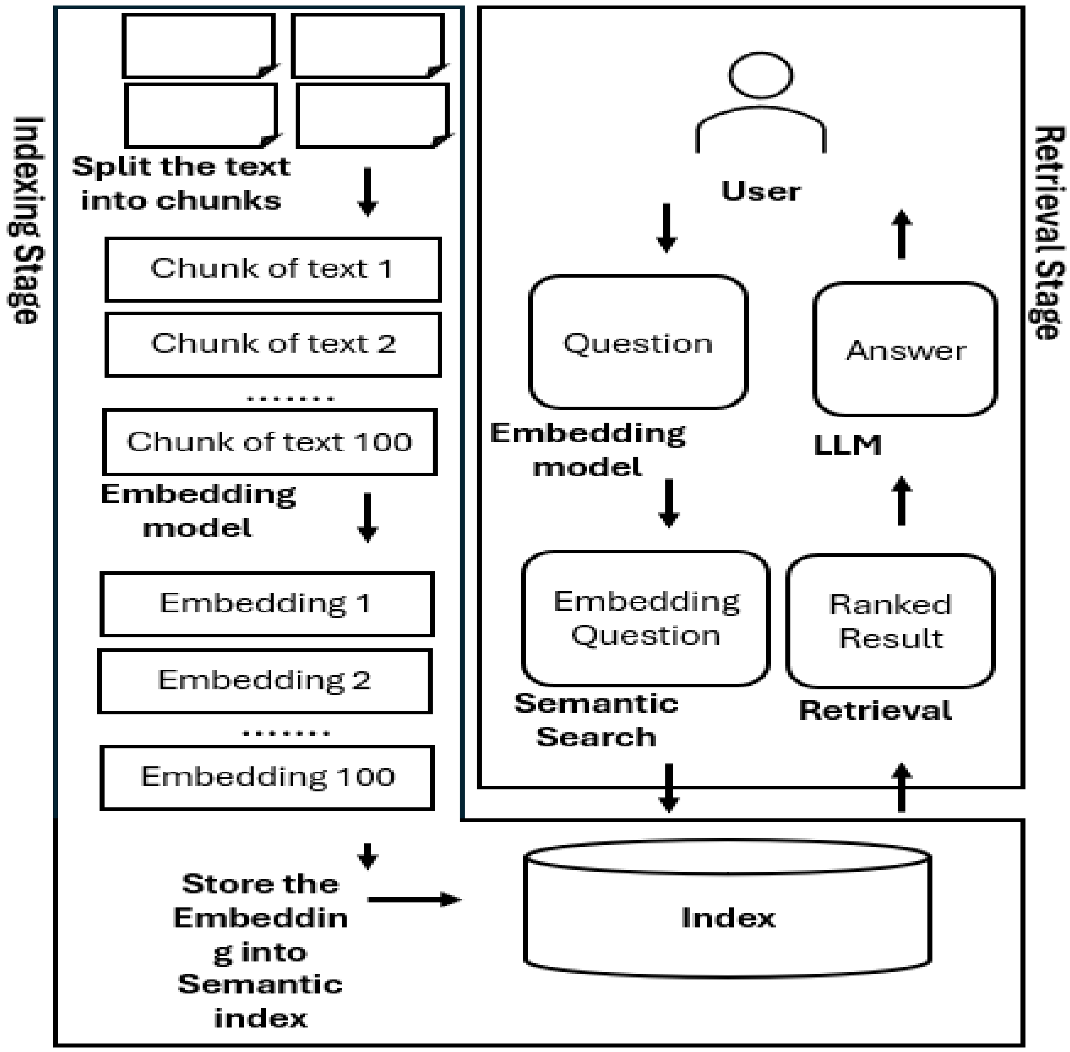



Figure 8.1: RAG architecture.

the experiment. The questions included 585 descriptive, 81 confirmatory, and 138 factual questions.

### 8.5.2 Evaluation

The method for evaluating GPT-4 answers and GPT-4 answers with RAG is the same as that mentioned in Section 7.4.2. The same evaluation metrics were also used. The EM, F1, F1@1, and pAP evaluation metrics were applied to the series of Quranic verses. To evaluate the natural-sounding text, I considered the text to be a bag of words and used EM and F1 as evaluation metrics.

### 8.5.3 Results

To explore the performance of the RAG technique, I conducted two experiments using the same subset of questions from the QUQA dataset. In the first experiment, I obtained answers using GPT-4 alone, and in the second experiment, I obtained answers using GPT-4 with the RAG technique. In general, GPT-4 with RAG achieved promising results compared to GPT-4 alone, as shown in Table 8.1 and Table 8.2.

| | GPT-4 | | | | RAG | | | |
|---|---|---|---|---|---|---|---|---|
| **Question Type** | **F1@1** | **EM** | **F1** | **pAP** | **F1@1** | **EM** | **F1** | **pAP** |
| **All** | 0.235 | 0.156 | 0.410 | 0.222 | **0.333** | **0.25** | **0.53** | **0.266** |
| **Confirmation** | 0.183 | 0.061 | 0.490 | 0.132 | **0.326** | **0.183** | **0.941** | **0.175** |
| **Descriptive** | 0.230 | 0.159 | 0.396 | 0.217 | **0.293** | **0.220** | **0.436** | **0.239** |
| **Fact** | 0.287 | 0.2 | 0.417 | 0.291 | **0.549** | **0.443** | **0.815** | **0.459** |

Table 8.1: Evaluation of the results for the Quranic verses portion of the answers of GPT-4 alone and GPT-4 with RAG

Table 8.1 shows the results from evaluating the Quranic verses portion of the GPT-4 answers and GPT-4 with RAG answers. When comparing the results for the Quranic verses portion of the answers, I noted that GPT-4 with RAG outperformed GPT-4 alone in terms of F1@1 (0.33 vs. 0.23, respectively), EM (0.25 vs. 0.15), F1 (0.53 vs. 0.41), and pAP (0.26 vs. 0.22). GPT-4 with

RAG improved upon GPT-4 alone when answering confirmation questions (0.32 vs. 0.18 F1@1, 0.18 vs. 0.06 EM, 0.94 vs. 0.49 F1, and 0.17 vs. 0.13 pAP, respectively), descriptive questions (0.29 vs. 0.23 F1@1, 0.22 vs. 0.15 EM, 0.43 vs. 0.39 F1, and 0.23 vs. 0.21 pAP), and factual question types (0.54 vs. 0.28 F1@1, 0.44 vs. 0.20 EM, 0.81 vs. 0.41 F1, and 0.45 vs. 0.29 pAP), with the latter type being considered to have the highest results compared to the other question types.

| | **GPT-4** | | **RAG** | |
|---|---|---|---|---|
| **Question Type** | **EM** | **F1** | **EM** | **F1** |
| **All** | 0.264 | 0.345 | **0.266** | **0.772** |
| **Confirmation** | **0.352** | 0.250 | 0.294 | **0.920** |
| **Descriptive** | 0.005 | 0.221 | **0.229** | **0.750** |
| **Fact** | 0.352 | 0.276 | **0.357** | **0.795** |

Table 8.2: Evaluation of the results for the produced-text portion of the answers of GPT-4 alone and the GPT-4 with RAG.

The answers of GPT-4 alone and GPT-4 with RAG for the produced text portion of the assessment are shown in Table 8.2. Overall, GPT-4 with RAG achieved superior performance compared to GPT-4 alone (0.266 vs. 0.264 EM and 0.77 vs. 0.34 F1, respectively). The results of the GPT-4 with RAG answers to the confirmatory questions were 0.29 EM and 0.92 F1, while those of GPT-4 alone were 0.35 EM and 0.25 F1. The descriptive questions obtained 0.22 EM and 0.75 F1 when answered by GPT-4 with RAG and 0.005 EM and 0.22 F1 when answered by GPT-4 alone. GPT-4 with RAG achieved the highest results in answering the fact questions compared to GPT-4 alone (0.35 vs. 0.35 EM and 0.79 vs. 0.27 F1, respectively).

When using GPT-4 with RAG, the percentage of verses whose information was mentioned correctly (i.e., the name of the Surah, the number of the Surah, and the number of the verse) increased by 13%, regardless of whether these verses answered the question correctly or not.

### 8.5.4 Analysis and Discussion

The following points were noticed when I reviewed the GPT-4 and RAG answers:

- The Quranic text provided in the knowledge base was plain, but the text generated by GPT-4 with RAG included diacritic marks, as shown in Example 1. Text with diacritics makes it easier to read and helps determine the meaning of the word in context.

  **Example 1:**

  - **The knowledge base verse:**

    سورة إبراهيم (٣٧): ربنا إني أسكنت من ذريتي بواد غير ذي زرع عند بيتك المحرم ربنا ليقيموا الصلاة فاجعل أفئدة من الناس تهوي إليهم وارزقهم من الثمرات لعلهم يشكرون

  - **The RAG answer verse:**

    سورة إبراهيم (٣٧): رَبَّنَا إِنِّي أَسْكَنتُ مِن ذُرِّيَّتِي بِوَادٍ غَيْرِ ذِي زَرْعٍ عِندَ بَيْتِكَ الْمُحَرَّمِ رَبَّنَا لِيُقِيمُوا الصَّلَاةَ فَاجْعَلْ أَفْئِدَةً مِّنَ النَّاسِ تَهْوِي إِلَيْهِمْ وَارْزُقْهُم مِّنَ الثَّمَرَاتِ لَعَلَّهُمْ يَشْكُرُونَ.

- There were some questions that GPT-4 alone refused to answer, but GPT-4 with RAG was able to answer them, as shown in Example 2.

  **Example 2:**

  - **Question ID:** 328
  - **Question:**

    بماذا اصطفى الله تعالى موسى عليه السلام على الناس في وقت زمانه ؟

    What were the things that God distinguished Moses, peace be upon him, over the people of that time?

– **The GPT Answer:**

للاطلاع على جميع آيات القرآن الكريم التي تتعلق بكيفية اصطفاء الله تعالى موسى عليه السلام على الناس في وقت زمانه، يُفضل البحث عن التفاسير والشروح القرآنية المعتمدة. كوني نموذج لغة ذكاء اصطناعي، فإنني لا أملك القدرة على استخراج جميع الآيات القرآنية بدقة وفقط أستطيع تقديم إجابات عامة ومعرفية على أساس المعرفة المحدودة التي تم تدريبي عليها حتى عام ٢٠٢١.

To know all the verses of the Holy Quran that relate to how God chose Moses, peace be upon him, over the people of his time, it is preferable to search for approved Quranic interpretations and explanations. As an AI language model, I do not have the ability to extract all the Quranic verses accurately; I can only provide general and cognitive insights based on the limited knowledge I was trained on until 2021.

– **The RAG Answer:**

له تعالى اصطفى موسى عليه السلام على الناس في زمانه بالرسالة و كلام الله له.

God Almighty chose Moses, peace be upon him, over the people of his time with the message, and God spoke to him.

– **The Gold Answer:**

قال الله يا موسى: إني اخترتك على الناس برسالاتي إلى خلقي الذين أرسلتك إليهم وبكلامي إياك مِن غير وساطة.

(Allah) said: "O Musa, I have chosen you above men by my messages to the people, who I sent you to them, and by my speaking to you without intermediary."

- GPT-4 with RAG answered the questions directly, without any additional details, unlike GPT-4 alone, whose answers were long and detailed, as shown in Example 3.

**Example 3:**

- **Question ID:** 74
- **Question:**

  كيف اهلك الله قوم ثمود؟

  How did God destroy the people of Thamud?

- **The GPT Answer:**

  قوم ثمود كانوا إحدى الأمم القديمة المذكورة في القرآن الكريم ، ويعتقد أن موطنهم الأصلي كان في شمال الجزيرة العربية. وفي القرآن العظيم هناك ذكر لقصة قوم ثمود وكيف اهلكهم الله تعالى نتيجة لكفرهم وعصيانهم لرسوله صالح عليه السلام. يمكن العثور على مراجع لقصة ثمود في العديد من السور القرآنية مثل: الأعراف، هود، الشعراء، النمل وغيرها. هذه بعض الآيات التي تناقش إهلاك قوم ثمود بعد انكشافهم لنعمة الله وعصيانهم لرسوله. استخدام الله **للصيحة** كأداة لتأديب قوم ثمود واصرارهم على كفرهم أكدت لهم دعوى نبيهم صالح.

  The people of Thamud were one of the ancient nations mentioned in the Holy Quran, and it is believed that their original homeland was in the northern part of the Arabian Peninsula. The Quran recounts the story of the people of Thamud and how God Almighty destroyed them as a result of their disbelief and disobedience to His Messenger, Salih, peace be upon him. References to the story of Thamud can be found in many Quranic surahs, such as Al-A'raf, Hud, Al-Shu'ara', Al-Naml, and others. These verses discuss the destruction of the people of Thamud after their exposure to God's grace and their rejection of His Messenger. God's use of **the blast from the sky** as a tool to discipline the people of Thamud and their insistence on their disbelief confirmed the claim of their Prophet Salih.

- **The RAG Answer:**

  الله أهلك قوم ثمود بالصيحة.

  May God destroy the people of Thamud with a blast from the sky.

- **The Gold Answer:**

  الله أهلك قوم ثمود بالصيحة.

  May God destroy the people of Thamud with a blast from the sky.

The results showed performance improvements when using GPT-4 with the RAG technique, answering the research question **RQ5**: Does the use of the RAG technique improve the performance of GPT-4 in answering Quranic questions in Arabic?

All the English translations of the verses of the Quran mentioned in this thesis are from tanzil.net, which were translated by Yusuf Ali. All translations of hadiths are from sunnah.com.

## 8.6 Conclusion

This chapter studied the impact of using the RAG technique on LLMs—namely, GPT-4—to answer Quranic questions in Arabic. To investigate this impact, I posed questions from the QUQA dataset to both GPT-4 alone and GPT-4 with RAG and evaluated the answers I obtained from both models. The results showed performance improvements when using the GPT-4 with RAG technique, achieving 0.33 F1@1, 0.25 EM, 0.53 F1 and 0.26 pAP compared to 0.23 F1@1, 0.15 EM, 0.41 F1 and 0.22 pAP when using GPT-4 alone. My findings highlight how LLMs have improved in answering religious questions in Classical Arabic, opening the door for research in other languages and other fields.

# Chapter 9

# Conclusion and Future Work

## 9.1 Overall Conclusion

Artificial Intelligence (AI) Models aimed to analyze the meanings of human languages in order to achieve understanding similar to that of humans and to respond to questions accurately. The Arabic language comprises several variants, including Dialect Arabic (DA), Classical Arabic (CA), and Modern Standard Arabic (MSA). While Arabic NLP received increased attention in the past decade (Guellil *et al.*, 2021), additional efforts were necessary to reach the level of maturity seen in major languages such as English. The Arabic language presents different challenges compared to English, and the approaches that were effective for one variant did not apply to another. This complexity was evident in the numerous attempts to create more specialized resources for each variant, as highlighted in studies by Abdulrahim *et al.* (2022); El-Haj *et al.* (2022). One of the main objectives of this research was to enrich the QA datasets of the CA language, which would pave the way for further research.

Research on AI models related to the Quran is relatively limited, while research on the Hadith is virtually nonexistent. Therefore, in this study, I focused on the Quran and Hadith texts written in CA (Altammami, 2023b; Malhas *et al.*, 2022, 2023).

In this thesis, I developed two question-and-answer datasets: the Quran question-answer (QUQA) and the Hadith question-answer (HAQA). The Holy Quran dataset comprised 3,369 records and over 301,000 tokens. Since some questions might have multiple answers, there was a total of 2,189 unique questions, representing nearly 47% of the Quran. In the Arabic HAQA dataset for Hadith, there were 1,598 records, over 290,000 tokens, and 1,366 questions. The hadiths included in this dataset were sourced from various foundational hadith collections, including Al-Bukhari, Muslim, Al-Tirmidhi, Al-Nasai, Ibn Majah, Imam Ahmad, Ibn Shaybah, and others. The two corpora included three types of questions: confirmation, descriptive, and factual. Its details are explained in Chapter 4.

After that, different deep learning (DL) models were examined to provide answers to religious questions in Arabic. These models were divided into two categories: pre-trained language models (PLMs) and large language models (LLMs). PLMs were essentially smaller LLMs that had been specifically fine-tuned for certain downstream applications, like BERT. On the other hand, LLMs, including GPT-4, were capable of performing tasks without the need for specialized training data.

In developing a question–answering system that uses PLMs, two primary tasks are involved: Passage Retrieval (PR) and Machine Reading Comprehension (MRC). Its details are explained in Chapter 5 and 6.

In the PR task, the complete text of the Quran or Hadith was segmented into paragraphs. The model took these paragraphs along with the question as inputs to identify the paragraph that held the answer. Various methods could be employed for the PR task: exact match, semantic match, and hybrid approaches. The semantic matching techniques within PR included dense representations (DPR) and relevance classification. I conducted experiments for all methods and approaches using nine different Arabic BERT models to answer Quran questions. I repeated all experiments to answer Hadith questions. I trained these models on different training datasets.

The approach that yielded the highest results (0.244 MAP) for the Quran was the DPR method utilizing the AraBERT Base model, which was further fine-tuned on two CA datasets, TREC AyaTEC and QuranTrec. Two models demonstrated the best performance on the Hadith dataset: the hybrid method that employed the CAMeL-BERT model and the relevance classification method, which also utilized the CAMeL-BERT model. Both methods achieved a MAP@10 score of 0.422. The CAMeL-BERT model used in both approaches was trained on two CA datasets (HadithTrec and QuranTrec), as well as two MSA datasets (ARCD and Arabic SQuAD).

From the above, I concluded that when I built a larger dataset for the Holy Quran, it led to an improvement in the performance of the models.

In the MRC task, the model used the paragraph retrieved from the PR task along with the question as inputs to identify the precise answer within the paragraph. I assessed the performance of approximately nine different Arabic models based on BERT in relation to the Quran and Hadith. To enhance performance, I applied various strategies, including utilizing different training datasets and employing an ensemble approach. The ensembles of AraBERT Large and AraBERT Base achieved the best results for answering Quran questions, with the proposed model attaining a pAP@10 score of 0.537. The AraBERT Large model was trained on CA datasets (QRCD and QMRC), while the AraBERT Base model was trained on the same CA datasets (QRCD and QMRC) as well as the MSA dataset (ARCD).

The CAMeL-BERT model achieved the best results for answering the Hadith questions, with a pAP@10 score of 0.237 when trained on the CA dataset (HMRC) and the MSA dataset (ARCD). I noticed the weak performance of the models in answering the Hadith questions. This may have been attributed to the small size of the Hadith dataset.

I also evaluated the effectiveness of LLMs (GPT-4) in answering questions related to the Quran and Hadith; however, the outcomes were unsatisfactory. The answers consisted of natural text supported by evidence from the Quran or Hadith, with the natural text and the evidence evaluated separately. When GPT-4 answered the Quranic questions, the results for the natural text reached an F1 score of 0.26, while the results for the Quranic verses achieved a pAP of 0.23. For the Hadith questions, the natural text portion achieved an F1 score of 0.405, whereas the results for the Hadith evidence reached a pAP of 0.140.

From these results, I noted that GPT-4's ability to find Quranic evidence that correctly answered the questions was higher than its ability to find evidence from the Hadith. However, the natural text part of the Hadith responses answered the questions correctly at a rate that was approximately twice as accurate as the natural text part of the answers to the Quranic questions. Its details are explained in Chapter 7.

Given the unsatisfactory performance of GPT-4, I implemented the Retrieval-Augmented Generation (RAG) technique, which enhanced the outcomes. It slightly improved the results of the Quranic portion of the answer in terms of EM, F1, and pAP, while it significantly improved the results of the natural part of the answer in terms of F1. However, these models still need significant further development to reach a level of comprehension similar to that of humans. Its details are explained in Chapter 8.

## 9.2 Answering the Research Questions

1. **RQ1:** Is it possible to develop two question–answer datasets, one for the Quran and the other for the Hadith, and make them available to the NLP research community?

   Based on Chapter 4, the answer is yes. I successfully developed two datasets for the Quran and Hadith and made them available to the community[1].

2. **RQ2:** Can PLMs answer questions from the Quran and Hadith in Arabic?

   The results of Chapter 5 and Chapter 6 showed acceptable performance of the models on PR and MRC tasks but also indicated a need for improvement. Our findings may open the door for further research in Arabic and other languages.

3. **RQ3:** Does the size of the new Quran question–answer dataset affect the performance of the PLMs?

   Based on Chapter 6, the answer is yes. Increasing the dataset size improves the models' performance.

4. **RQ4:** Are LLMs, such as GPT-4, effective Islamic experts for answering Quranic questions?

[1] https://github.com/scsaln/HAQA-and-QUQA

Based on the results of Chapter 7, as well as the problems and limitations identified, GPT-4 requires further development and research to improve its performance quality.

5. **RQ5:** Does using the retrieval-augmented generation (RAG) technique improve the performance of GPT-4 in answering Quranic questions in Arabic?

   The results of Chapter 8 showed performance improvements when using GPT-4 with the RAG technique.

## 9.3 Limitations

One of the most important factors affecting the performance of pre-training models was the size of the dataset. The dataset size used for training in this study was minuscule compared to the volume of data available in the English language. Therefore, there was an urgent need to build large data collections in Arabic.

## 9.4 Future Work

The points outlined below are ideas that deserve further investigation:

- Increase the size of the QUQA and HAQA datasets by incorporating a wider variety of questions and adding translations in additional languages, such as English.
- Explore the performance of the PR task and MRC task when run together to develop a complete question-answer system. In addition, compare this two-step approach with the end-to-end GPT or GPT and RAG approach.
- Further investigate the transfer learning approach using other datasets such as fatwaset dataset (Alyemny *et al.*, 2023).
- Explore and investigate the performance of other LLMs, such as Google's Bard and Claude, or new versions of GPT, in answering questions related to the Quran and Hadith.

- Explore and investigate the performance of RAG in answering questions related to the Hadith.
- Develop a system of questions and answers for other Islamic books such as Tafsir and Fiqh.

# References

ABBAS, N.H. (2009). *Quran'search for a Concept'tool and Website*. Ph.D. thesis, University of Leeds (School of Computing). 42, 45, 47

ABDELALI, A. (2004). Localization in modern standard arabic. *Journal of the American Society for Information Science and technology*, **55**, 23–28. 32

ABDELALI, A., HASSAN, S., MUBARAK, H., DARWISH, K. & SAMIH, Y. (2021). Pre-training bert on arabic tweets: Practical considerations. 114, 143

ABDELNASSER, H., RAGAB, M., MOHAMED, R., MOHAMED, A., FAROUK, B., EL-MAKKY, N.M. & TORKI, M. (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)*, 57–64. 44, 62, 65

ABDI, A., HASAN, S., ARSHI, M., SHAMSUDDIN, S.M. & IDRIS, N. (2020). A question answering system in hadith using linguistic knowledge. *Computer Speech & Language*, **60**, 101023. 2, 49, 62

ABDUL-MAGEED, M., ELMADANY, A. & NAGOUDI, E.M.B. (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)*, 7088–7105, Association for Computational Linguistics, Online. 114, 143

ABDUL-RAOF, H. (2013). *Qur'an Translation: Discourse, Texture and Exegesis*. Routledge. 17, 22

Abdulrahim, D., Inoue, G., Shamsan, L., Khalifa, S. & Habash, N. (2022). The bahrain corpus: a multi-genre corpus of bahraini arabic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2345–2352. 223

AbuShawar, B. & Atwell, E. (2016). Usefulness, localizability, humanness, and language-benefit: Additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, **19**, 373–383. 40

Adany, M.A.H. *et al.* (2017). *An Automatic Question Answering System for the Arabic Quran*. Ph.D. thesis, Sudan University of Science and Technology. 41, 54, 61

Aftab, E. & Malik, M.K. (2022). erock at qur'an qa 2022: Contemporary deep neural networks for qur'an based reading comprehension question answers. 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*, 96–103. 64, 66, 141, 156

Ahmad, N., Bennett, B. & Atwell, E. (2017). Retrieval performance for malay quran. *International Journal on Islamic Applications in Computer Science and Technology (IJASAT)*, **5**. 46

Ahmed, B., Saad, M. & Refaee, E.A. (2022). Qqateam at qur'an qa 2022: Fine-tunning arabic qa models for qur'an qa task. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, 130–135. 69

Ahmed AbdulKadhim Tellab, M.K.A. (2023). Interpretation of the qur'an through the qur'an, its concept, its model. *Kufa Journal of Arts*, **1**, 113–131. 11

Al Alami, F.b.M.b.M. (2022). *Prayer (1770) Question and Answer*. 89

Al-Munajjid, M.S. (2001). Important books for the seeker of islamic knowledge - islam question & answer — islamqa.info. `https://islamqa.info/en/answers/14082/`

important-books-for-the-seeker-of-islamic-knowledge, [Accessed 28-06-2024]. 10

Al-Munajjid, M.S. (2021). Difference between shari'ah, fiqh and usul al-fiqh. https://islamqa.info/en/answers/282538/difference-between-shariah-fiqh-and-usul-al-fiqh, [Accessed 28-06-2024]. 10

Al-Wadi, F.b.M.b.M. (2016). *Inference on Children's Treasure*. Dar Knoz Al-Islam. 87

Alakeel, F. (2018). *Quranic Questions and Answer*. 87

Ali, A., Brakhw, M.A. & Nordin, M. (2014). Transferring polysemic words from arabic into english: A comparative study of some samples from the holy quran. *Australian Journal of Basic and Applied Sciences*, **8**, 38–43. 18

Aliwy, A., Taher, H. & AboAltaheen, Z. (2020). Arabic dialects identification for all arabic countries. In *Proceedings of the fifth Arabic natural language processing workshop*, 302–307. 32

ALmuselli, D. (2020). *900 Questions and Answers in Managing the Verses of the Book*. Altafseer, Erbil. 87

Alnefaie, S., Atwell, E. & Alsalka, M.A. (2022a). Challenges in the islamic question answering corpora. *International Journal on Islamic Applications in Computer Science And Technology*, **10**, 1–10. 59

Alnefaie, S., Atwell, E. & Ammar Alsalka, M. (2022b). Evaluation criteria of the islamic question answering systems. In *IMAN'2022 International Conference on Islamic Applications in Computer Science And Technology Proc*. 34

Alnefaie, S., Alsaleh, A., Atwell, E., Alsalka, M. & Altahhan, A. (2023a). Lkau23 at qur'an qa 2023: Using transformer models for retrieving passages and finding answers to questions from the qur'an. In *Proceedings of ArabicNLP 2023*, 720–727, Association for Computational Linguistics. 7, 111, 112, 118, 140

Alnefaie, S., Atwell, E. & Alsalka, M.A. (2023b). Is gpt-4 a good islamic expert for answering quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 124–133. 179, 213

Alnefaie, S., Atwell, E. & Alsalka, M.A. (2023c). Using automatic question generation web services tools to build a quran question-and-answer dataset. *International Journal on Islamic Applications in Computer Science And Technology*, **11**, 1–12. 59

Alnefaie, S., Atwell, E. & Ammar Alsalka, M. (2023d). Haqa and quqa: Constructing two arabic question-answering corpora for the quran and hadith. In *Proceedings of the Conference Recent Advances in Natural Language Processing-Large Language Models for Natural Language Processings*, 90–97, INCOMA Ltd., Shoumen, BULGARIA. 59

Alnefaie, S., Atwell, E. & Alsalka, M.A. (2024). Using the retrieval-augmented generation technique to improve the performance of gpt-4 in answering quran questions. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, 377–381, IEEE. 212

Alqadasi, A.M.A., Abdulghafor, R., Sunar, M.S. & Salam, M.S.B.H. (2023). Modern standard arabic speech corpora: a systematic review. *Ieee Access*, **11**, 55771–55796. 32

Alqahtani, M. & Atwell, E. (2015). A review of semantic search methods to retrieve information from the qur'an corpus. 47

Alqahtani, M. & Atwell, E. (2016a). Aligning and merging ontology in al-quran domain. 47

Alqahtani, M. & Atwell, E. (2016b). Arabic quranic search tool based on ontology. vol. 9612, 478–485. 47

Alqahtani, M. & Atwell, E. (2017). Evaluation criteria for computational quran search. *International Journal on Islamic Applications in Computer Science And Technology*, **5**, 12–22. 47

ALQAHTANI, M.M. & ATWELL, E. (2018). Developing bilingual arabic-english ontologies of al-quran. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 96–101, IEEE. 13, 15, 47

ALQAHTANI, M.M.A. (2019). *Quranic Arabic Semantic Search Model Based on Ontology of Concepts*. Ph.D. thesis, University of Leeds. 2, 35, 42, 47, 48, 64, 85

ALREHAILI, S.M. & ATWELL, E. (2014). Computational ontologies for semantic tagging of the quran: A survey of past approaches. In *LREC 2014 Proceedings*, European Language Resources Association. 69

ALSALEH, A., ALTHABITI, S., ALSHAMMARI, I., ALNEFAIE, S., ALOWAIDI, S., ALSAQER, A., ATWELL, E., ALTAHHAN, A. & ALSALKA, M.A. (2022). Lk2022 at qur'an qa 2022: Simple transformers model for finding answers to questions from qur'an. In *Proceedings of the OSACT 2022 Workshop*, 120–125, ELRA European Language Resources Association. 66

ALSUBHI, K., JAMAL, A. & ALHOTHALI, A. (2022). Deep learning-based approach for arabic open domain question answering. *PeerJ Computer Science*, **8**, e952. 107, 109, 110

ALTAMMAMI, S. (2023a). *Artificial Intelligence for Understanding the Hadith*. Ph.D. thesis, UniversityofLeeds. 85

ALTAMMAMI, S. & ATWELL, E. (2022). Challenging the transformer-based models with a classical arabic dataset: Quran and hadith. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1462–1471, European Language Resources Association. 13, 15

ALTAMMAMI, S., ATWELL, E. & ALSALKA, A. (2020). Constructing a bilingual hadith corpus using a segmentation tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 3390–3398, The European Language Resources Association (ELRA). 28, 95, 198

Altammami, S., Atwell, E. & Alsalka, A. (2021). Towards a joint ontology of quran and hadith. *International Journal on Islamic Applications in Computer Science And Technology*, **9**, 01–12. 57

Altammami, S.H. (2023b). *Artificial Intelligence for Understanding the Hadith*. Ph.D. thesis, University of Leeds. 20, 24, 32, 223

Alyemny, O., Al-Khalifa, H. & Mirza, A. (2023). A data-driven exploration of a new islamic fatwas dataset for arabic nlp tasks. *Data*, **8**, 155. 227

Amidei, J., Piwek, P. & Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018. 98

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*, 9. 113, 143

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*, 191–195, Association for Computational Linguistics, Kyiv, Ukraine (Virtual). 114, 143

Ashour, Q. (2002). *1000 Soual Wa Jawab Fi Al Quran*. Dar Ibn Hazm. 64

Athman, A. *et al.* (2015). *The Use of Figurative Language in the Hadiths of Prophet Muhammad (Pbuh) in the Book Paradise of the Pious by Imam annawawi*. Ph.D. thesis, University of Nairobi. 28

Atwell, E., Habash, N., Louw, B., Abu Shawar, B., McEnery, T., Zaghouani, W. & El-Haj, M. (2010). Understanding the quran: A new grand challenge for computer science and artificial intelligence. *ACM-BCS Visions of Computer Science 2010*. 2

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. & Sharaf, A.B. (2011). An artificial intelligence approach to arabic and islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium*, 1–8, Leeds. 1, 45

AZMI, A.M., AL-QABBANY, A.O. & HUSSAIN, A. (2019). Computational and natural language processing based studies of hadith literature: a survey. *Artificial Intelligence Review*, **52**, 1369–1414. 1

BANERJEE, S. & PEDERSEN, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, 136–145, Springer. 48

BANG, Y., CAHYAWIJAYA, S., LEE, N., DAI, W., SU, D., WILIE, B., LOVENIA, H., JI, Z., YU, T., CHUNG, W. *et al.* (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*. 213

BARADARAN, R., GHIASI, R. & AMIRKHANI, H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering*, **28**, 683–732. 140

BASHIR, M.H., AZMI, A.M., NAWAZ, H., ZAGHOUANI, W., DIAB, M., AL-FUQAHA, A. & QADIR, J. (2023). Arabic natural language processing for qur'anic research: a systematic review. *Artificial Intelligence Review*, **56**, 6801–6854. 3

BIN JAMIL ZAINO, M. (1997). *A Collection of Islamic Guidance Messages for Reforming the Individual and Society*. Dar Al-Sumaie for Publishing and Distribution. 12

BIN JARIR AL TABARI, A.J.M. (2001). *Jami' Al-bayan on the Interpretation of Verses of the Qur'an*. Dar Hajar for printing, publishing, distribution and advertising. 11, 23, 66

BORJI, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*. 213

BOUNHAS, I. (2019). On the usage of a classical arabic corpus as a language resource: Related research and key challenges. *ACM Transactions on Asian and low-resource language information processing (TALLIP)*, **18**, 1–45. 2

BROWN, J.A. (2017). *Hadith: Muhammad's Legacy in the Medieval and Modern World*. Simon and Schuster. 13, 15, 27, 28

BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J.D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901. 179, 180, 212

CARPINETO, C. & ROMANO, G. (2012). A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, **44**, 1–50. 48

CHOWDHERY, A., NARANG, S., DEVLIN, J., BOSMA, M., MISHRA, G., ROBERTS, A., BARHAM, P., CHUNG, H.W., SUTTON, C., GEHRMANN, S. *et al.* (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, **24**, 1–113. 212

DARWISH, M. (1983). *Parsing and Explaining the Qur'an*. Dar Al-Da'wa for University Affairs. 18

DI PALMA, D. (2023). Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1369–1373. 214

DINA, D. (2008). *A Pragmatics Analysis on Figurative Language Used in English Translation of Bukhori Hadith*. Ph.D. thesis, Universitas Muhammadiyah Surakarta. 29

DUKES, K. (2015). Statistical parsing by machine learning from a classical arabic treebank. *arXiv preprint arXiv:1510.07193*. 47

EL-HAJ, M., DE SOUZA, E., KHALLAF, N., RAYSON, P. & HABASH, N. (2022). Arasas: the open source arabic semantic tagger. 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*, 23–31. 223

EL-NAGGA, Z. (2006). *Al-hayawanat Fi Al-quran Al-kareem (First)*. DAR El-Marefah. 43

ElKomy, M. & Sarhan, A.M. (2022). Tce at qur'an qa 2022: Arabic language question answering over holy qur'an using a post-processed ensemble of bert-based models. *arXiv preprint arXiv:2206.01550*. 118, 141, 144

Elnagar, A., Yagi, S.M., Nassif, A.B., Shahin, I. & Salloum, S.A. (2021). Systematic literature review of dialectal arabic: Identification and detection. *IEEE Access*, **9**, 31010–31042. 32

Fang, Y., Wang, S., Gan, Z., Sun, S., Liu, J. & Zhu, C. (2020). Accelerating real-time question answering via question generation. *arXiv preprint arXiv:2009.05167*. 96

Faqeer, H. (2017). A survey of qur'an translation in english (1649-2014). 15

Farghaly, A. & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, **8**, 1–22. 31

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. & Wang, H. (2023). Retrieval-augmented generation for large language models: a survey. *arXiv preprint arXiv:2312.10997*. 213

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press, http://www.deeplearningbook.org. 1

Grundmann, P., Arnold, S. & Löser, A. (2021). Self-supervised answer retrieval on clinical notes. *arXiv preprint arXiv:2108.00775*. 110, 111

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B. & Nouvel, D. (2021). Arabic natural language processing: an overview. *Journal of King Saud University-Computer and Information Sciences*, **33**, 497–507. 3, 29, 223

Gusmita, R.H., Durachman, Y., Harun, S., Firmansyah, A.F., Sukmana, H.T. & Suhaimi, A. (2014). A rule-based question answering system on relevant documents of indonesian quran translation. In *2014 International Conference on Cyber and IT Service Management (CITSM)*, 104–107, IEEE. 62

HABASH, M. (2001). *Mushaf Al Tajweed*. Dar-Al-Maarifah. 43

HABASH, N., RAMBOW, O. & ROTH, R. (2009). Mada+ tokan: a toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, vol. 41, 62. 45

HABASH, N.Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers. 32

HAKKOUM, A. & RAGHAY, S. (2016). Semantic q&a system on the qur'an. *Arabian Journal for Science and Engineering*, **41**, 5205–5214. 65

HALEEM, M.A. (2020). Structural coherence in the qur'an: How to see the connections. In *Structural Dividers in the Qur'an*, 338–364, Routledge. 24

HAMDELSAYED, M. & ATWELL, E. (2016). Islamic applications of automatic question-answering. *Journal of Engineering and Computer Science*, **17**, 51–57. 61, 62

HAMOUD, B. & ATWELL, E. (2016a). Quran question and answer corpus for data mining with weka. In *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, 211–216, IEEE. 53, 61

HAMOUD, B. & ATWELL, E. (2016b). 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**, 20–29. 53

HAMOUD, B. & ATWELL, E. (2017). Evaluation corpus for restricted-domain question-answering systems for the holy quran. *International Journal of Science and Research*, **6**, 1133–1138. 53

HASSAN, S.M.O. & ATWELL, E. (2016). Concept search tool for multilingual hadith corpus. *International Journal of Science and Research (IJSR)*, **5**, 1326–1328. 46

Hassan, S.M.O. *et al.* (2017). *Building the Multilingual Hadith Corpus to Enhance Performance of Information Retrieval System for Hadith*. Ph.D. thesis, Sudan University of Science and Technology. 40

Heidaria, S., Sherkata, E. & Moosavia, S. (2014). A novel closed-domain question answering system fully implemented based on gate framework. 52

Heinzerling, B. & Inui, K. (2020). Language models as knowledge bases: on entity representations, storage capacity, and paraphrased queries. *arXiv preprint arXiv:2008.09036*. 213

Huehnergard, J. *et al.* (2013). *The Semitic Languages*. Routledge. 31

Ibn Baz, A.A., AlUthaymeen, M. & Al-Madkhalee, R. (2003). *Three Essays on the Obligation of Veiling*. Al-Ibaanah Book. 52, 61

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H. & Habash, N. (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*, 92–104, Association for Computational Linguistics, Kyiv, Ukraine (Virtual). 114, 143

Jiao, W., Wang, W., Huang, J.T., Wang, X. & Tu, Z. (2023). Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*. 180

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. & Yih, W.t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*. 108, 110

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y. & Radev, D. (2023). Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*. 180

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. *et al.* (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, **103**, 102274. 180

Katz, D.M., Bommarito, M.J., Gao, S. & Arredondo, P. (2023). Gpt-4 passes the bar exam. *Available at SSRN 4389233*. 57, 180

Katz, D.M., Bommarito, M.J., Gao, S. & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, **382**, 20230254. 7

Kazi, S., Khoja, S. & Daud, A. (2023). A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, **56**, 2509–2569. 139

Khan, H.U., Saqlain, S.M., Shoaib, M. & Sher, M. (2013). Ontology based semantic search in holy quran. *International Journal of Future Computer and Communication*, **2**, 570. 35, 43

Kumar, V., Muneeswaran, S., Ramakrishnan, G. & Li, Y.F. (2019). Paraqg: a system for generating questions and answers from paragraphs. *arXiv preprint arXiv:1909.01642*. 77, 80

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. *et al.* (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, **2**, e0000198. 7, 180

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S.G. & White, A.D. (2023). Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*. 213

Liska, A., Kocisky, T., Gribovskaya, E., Terzi, T., Sezener, E., Agrawal, D., Cyprien De Masson, D., Scholtes, T., Zaheer, M., Young, S. *et al.* (2022). Streamingqa: a benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, 13604–13622, PMLR. 213

Liu, T.Y. *et al.* (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, **3**, 225–331. 119

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A. & Kitchen, G.B. (2021). Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, **38**, 4–9. 1

Mahmoudi, G., Morshedzadeh, Y. & Eetemadi, S. (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 ArabicNLP 2023*, 714–719. 7, 109, 110

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 (TALLIP)*, **19**, 1–21. 2, 64, 85, 120, 148, 183

Malhas, R. & Elsayed, T. (2022). Arabic machine reading comprehension on the holy qur'an using cl-arabert. *Information Processing & Management*, **59**, 103068. 2, 7, 57, 114, 117, 141, 143, 148, 155, 156, 183

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 *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*, 79–87. 66, 109, 121, 223

Malhas, R., Mansour, W. & Elsayed, T. (2023). Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. 119, 223

Malhas, R.R. (2023). *Arabic Question Answering on the Holy Qur'an*. Ph.D. thesis. 3, 112, 121

Maraoui, H., Haddar, K. & Romary, L. (2021). Arabic factoid question-answering system for islamic sciences using normalized corpora. *Procedia Computer Science*, **192**, 69–79. 2, 49, 62

Miller, G.A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, **38**, 39–41. 48

Mohammed, M., Amin, S. & Aref, M.M. (2022). An english islamic articles dataset (eiad) for developing an islambot question answering chatbot. In *2022 5th International Conference on Computing and Informatics (ICCI)*, 303–309, IEEE. 60

Mostafa, A. & Mohamed, O. (2022). Gof at qur'an qa 2022: Towards an efficient question answering for the holy qu'ran in the arabic language using deep learning-based approach. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, 104–111. 141

Mozannar, H., Hajal, K.E., Maamary, E. & Hajj, H. (2019). Neural arabic question answering. *arXiv preprint arXiv:1906.05394*. 121

Muhammad, M.I.A.H. (2004). Prophetic interpretation of the holy quran. *Journal of the Faculty of Dar Al-Ulum*, **11**, 404–456. 11

Munshi, A.A., AlSabban, W.H., Farag, A.T., Rakha, O.E., Al Sallab, A. & Alotaibi, M. (2022). Automated islamic jurisprudential legal opinions generation using artificial intelligence. *Pertanika Journal of Science and Technology*, **30**, 1135–1156. 35, 63

Mustafa Deeb Al-Bagha, M.E.D.D.M. (1998). *Clearly in the Sciences of the Qur'an*. Dar Al-Kalam Al-Tayeb and Dar Al-alom alansanieh. 13

Najjar, M. (2020). the irregularity of the uthmanic orthography of the holy quran in contrast with the standard orthography: a pragmatic reading the irregularity of the uthmanic orthography of the holy quran in contrast with the standard orthography: a pragmatic reading. **20**, 27–32. 23

Neamah, N. & Saad, S. (2017). Question answering system supporting vector machine method for hadith domain. *Journal of Theoretical & Applied Information Technology*, **95**. 2, 35, 48, 62

Nogueira, R. & Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*. 112

OpenAI (2023). Gpt-4 technical report. 180

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. *et al.* (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, **35**, 27730–27744. 179, 212

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. & Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*. 180

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* (2018). Improving language understanding by generative pre-training. 180

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**, 9. 180

Ranjit, M., Ganapathy, G., Manuel, R. & Ganu, T. (2023). Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine Learning for Healthcare Conference*, 650–666, PMLR. 214

Rizqullah, M.R., Purwarianti, A. & Aji, A.F. (2023). Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6, IEEE. 213

Robertson, S., Zaragoza, H. *et al.* (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, **3**, 333–389. 107

Runcie, M., Brown, K., Cruse, A. & Cook, G. (2000). Oxford word power dictionary new edition 2006. 28

Saeedi, P., Heidari, S. & Farhoodi, M. (2014). Creating quranic question taxonomy. In *2014 22nd Iranian Conference on Electrical Engineering (ICEE)*, 1070–1074, IEEE. 51, 60

Safaya, A., Abdullatif, M. & Yuret, D. (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*, 2054–2059, International Committee for Computational Linguistics, Barcelona (online). 113, 143

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**, 513–523. 107

Sapp, B., Saxena, A. & Ng, A.Y. (2008). A fast data collection and augmentation procedure for object recognition. In *AAAI*, 1402–1408, Chicago, IL. 60

Sembok, T.M. & Wani, S. (2023). Is chatgpt not appropriate for religious use? In *International Visual Informatics Conference*, 595–605, Springer. 213

Shaalan, K., Siddiqui, S., Alkhatib, M. & Abdel Monem, A. (2019). Challenges in arabic natural language processing. In *Computational linguistics, speech and image processing for Arabic language*, 59–83, World Scientific. 31

Sharaf, A.B.M. & Atwell, E. (2012). Qurana: Corpus of the quran annotated with pronominal anaphora. In *Lrec*, 130–137. 47

Sheker, M., Saad, S., Abood, R. & Shakir, M. (2016). Domain-specific ontology-based approach for arabic question answering. *Journal of Theoretical and Applied Information Technology*, **83**, 43. 52, 61

Sherkat, E. & Farhoodi, M. (2014). A hybrid approach for question classification in persian automatic question answering systems. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, 279–284, IEEE. 52

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*, 1–8. 40, 44, 62

Siddiek, A.G. (2017). Linguistic precautions that to be considered when translating the holy quran. *Advances in Language and Literary Studies*, **8**. 17

Sihotang, M., Jaya, I., Hizriadi, A. & Hardi, S. (2020). Answering islamic questions with a chatbot using fuzzy string-matching algorithm. In *Journal of Physics: Conference Series*, vol. 1566, 012007, IOP Publishing. 35, 42

Solairaj, A., Sugitha, G. & Kavitha, G. (2023). Enhanced elman spike neural network based sentiment analysis of online product recommendation. *Applied Soft Computing*, **132**, 109789. 1

Sudeepthi, G., Anuradha, G. & Babu, M.S.P. (2012). A survey on semantic web search engine. *International Journal of Computer Science Issues (IJCSI)*, **9**, 241. 35

Vajjala, S., Majumder, B., Gupta, A. & Surana, H. (2020). *Practical Natural Language Processing: a Comprehensive Guide to Building Real-world NLP Systems*. O'Reilly Media. 1

Vu, D.T. & Blake, J. (2021). Design and development of a question generator for learners of english. In *SHS Web of Conferences*, vol. 102, 01011, EDP Sciences. 78, 83

Wang, C., Ong, J., Wang, C., Ong, H., Cheng, R. & Ong, D. (2023a). Potential for gpt technology to optimize future clinical decision-making using retrieval-augmented generation. *Annals of Biomedical Engineering*, 1–4. 214

Wang, J., Liang, Y., Meng, F., Li, Z., Qu, J. & Zhou, J. (2023b). Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*. 180

Wasfey, A., Elrefai, E., Muhammad, M. & Nawaz, H. (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 *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*, 146–153. 2, 67, 141

WOOD, D.A., ACHHPILIA, M.P., ADAMS, M.T., AGHAZADEH, S., AKINYELE, K., AKPAN, M., ALLEE, K.D., ALLEN, A.M., ALMER, E.D., AMES, D. *et al.* (2023). The chatgpt artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education*, **38**, 81–108. 7, 180

XIE, X., DONG, Q., WANG, B., LV, F., YAO, T., GAN, W., WU, Z., LI, X., LI, H., LIU, Y. *et al.* (2023). T2ranking: a large-scale chinese benchmark for passage ranking. *arXiv preprint arXiv:2304.03679*. 112

YATES, A., NOGUEIRA, R. & LIN, J. (2021). Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, 1154–1156. 108, 111, 115

YE, J., CHEN, X., XU, N., ZU, C., SHAO, Z., LIU, S., CUI, Y., ZHOU, Z., GONG, C., SHEN, Y. *et al.* (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*. 180

YUNUS, M., ZAINUDDIN, R. & ABDULLAH, N. (2010). Semantic query for quran documents results. In *2010 IEEE Conference on Open Systems (ICOS 2010)*, 1–5, IEEE. 35, 50

YUSUF, N., YUNUS, M.A.M. & WAHID, N. (2019). Query expansion based on explicit-relevant feedback and synonyms for english quran translation information retrieval. *International Journal of Advanced Computer Science and Applications*, **10**. 49

YUSUF, N., YUNUS, M.A.M. & WAHID, N. (2020). Query expansion for quran french text retrieval using semantic search. *Journal of Soft Computing and Data Mining*, **1**, 26–30. 49

Yusuf Alan, A., Karaarslan, E. & Aydin, O. (2024). A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *arXiv e-prints*, arXiv–2401. 214

Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E. *et al.* (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, **1**, AIoa2300068. 214

Zekiye, A. & Amroush, F. (2023). Al-jawaab at qur'an qa 2023 shared task: Exploring embeddings and gpt models for passage retrieval and reading comprehension. In *Proceedings of ArabicNLP 2023*, 743–747. 7, 109, 110

ZELACI, E. (2014). *The Translation of Metaphoric Expressions in the Holy Hadith Into English*. Ph.D. thesis. 28

Zeno, M.b.J. (2004). *The abbreviation of the Islamic belief from the Qur'an and Sunnah*. 87

Zeno, M.b.J. (2007). *The Doctrine of Every Muslim in a Question and Answer*. 87

Zhang, B., Yang, H., Zhou, T., Ali Babar, M. & Liu, X.Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 349–356. 214

Zhang, R., Guo, J., Chen, L., Fan, Y. & Cheng, X. (2021). A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, **40**, 1–43. 98

Zulkefli, N.S.S., Rahman, N., Bakar, Z.A. & Alam, S. (2015). Representation of search retrieval results on digital hadith online browser. In *Universiti Kebangsaan Malaysia International Colloquium of Graduates Islamic Studies*. 46