

# LLMs Based Approach for Quranic Question Answering

Zakia Saadaoui<sup>1,2</sup><sup>a</sup>, Ghassen Tlig<sup>3</sup><sup>b</sup> and Fethi Jarray<sup>1,2</sup><sup>c</sup>

<sup>1</sup>*LIMITIC Laboratory, UTM University, Tunisia*

<sup>2</sup>*Higher Institute of Computer Science of Medenine, Tunisia*

<sup>3</sup>*Ecole Supérieure d'Electronique de l'Ouest, Paris, France*

**Keywords:** Quranic QAS, LLM, Prompt Engineering.

**Abstract:** This paper addresses a prominent research gap in Quranic question answering, where current methodologies face challenges in capturing the nuanced aspects of inquiries related to the Quran. By presenting an innovative approach that utilizes Large Language Models (LLMs), including GPT4, Bart, and LLAMA, we aim to overcome these limitations, improving the clarity and precision of responses to Quranic queries. The evaluation of the proposed Quranic Question Answering System, using the F1 score metric, demonstrates encouraging results in comprehending and addressing various Quranic queries. Notably, the application of a chain of thought with Bart achieves an impressive F1-score of 95%. This research offers a distinctive perspective on Quranic question answering through the integration of LLMs.

## 1 INTRODUCTION: QA SYSTEMS


Question Answering (QA) systems play a pivotal role in information retrieval and can be broadly categorized into two main types: Open Domain and Closed Domain. In Open Domain QA, models operate without constraints to a specific domain, drawing information from diverse sources like books, the internet, Wikipedia, tables, graphs, and knowledge bases. On the other hand, Closed Domain QA systems are tailored for specific domains, such as legal or healthcare documents. Both Open Domain and Closed Domain QA systems can be further classified into open and closed book QA systems. Addressing the complexities of the Quran presents unique challenges for question answering systems. Existing methods often struggle to grasp the intricacies of Quranic language and context, leading to incomplete or inaccurate responses. This paper addresses this critical gap in Quranic AI, proposing innovative approaches for more comprehensive and insightful answers. The complexity of Quranic questions requires advanced natural language processing (NLP) techniques such as the utilization of Large Language Models (LLMs).


Large Language Models, such as GPT4, Bart, and LLAMA, have demonstrated exceptional capabilities in understanding and generating human-like text across various domains. However, their application to answering Quranic questions remains relatively unexplored.


In this paper, we present a novel approach that leverages LLMs to enhance the comprehension and accuracy of Quranic question answering. Specifically, we focus on GPT4 tailoring their capabilities to the unique challenges posed by Quranic text. By doing so, we aim to fill the existing research gap and contribute to the advancement of Quranic question answering systems. The proposed Quranic Question Answering System is rigorously evaluated using the F1-score metric, providing insights into its effectiveness in comprehending and addressing a diverse array of Quranic queries. Notably, employing a chain of thought with Bart achieves an exceptional F1-score of 95%, underscoring the potential of this approach.

### 1.1 Architecture of Arabic QAS

A Question Answering System (QAS) typically consists of three several components :question analysis, passage retrieval, and answer extraction. Question Analysis module analyzes the structure of the question, identifies the question type : factual, (e.g., where, when, who, what) or a taxonomy manually

<sup>a</sup> <https://orcid.org/0000-0001-8695-2034>

<sup>b</sup> <https://orcid.org/0000-0002-1911-9170>

<sup>c</sup> <https://orcid.org/0000-0002-5110-1173>

defined by linguistic experts. Based on the question analysis, this module generates a structured query or representation that can be used to retrieve relevant information from a knowledge source. A retrieval is a component that retrieves relevant passages or documents from a knowledge base or a corpus. An Answer Extraction this is module identifies potential answer candidates within the retrieved passages or documents. Answer Ranking and Selection: Once answer candidates are identified, they may be ranked based on their relevance and confidence scores. The top-ranked answers are selected and returned to the user. In some cases, A QAS needs a natural language Understanding and Generation module to refine the answer through additional processing such as summarization or paraphrasing to ensure clarity and coherence and to generate a final answer to be formatted and presented to the user in natural language, along with any relevant contextual information or sources. the progress in NLP techniques, particularly with the advent of deep learning and transformer-based models like BERT, GPT, and their variants, has significantly improved the accuracy and capabilities of QASs. These models have led to innovations in various components of the QAS architecture, it enables end-to-end formability. For question analysis, work is developing neural classifiers to determine question types. For question analysis, work is underway to develop neural classifiers to determine question types. For example, the CNN-based model and the LSTM-based model are used respectively to classify given questions, and both achieve competitive results. For document retrieval, methods based on dense representation have been proposed to solve the "term mismatch" problem, which has long hampered retrieval performance. Unlike traditional methods such as TF-IDF and BM25 , deep search methods learn to encode questions and documents in a latent vector space where text semantics beyond term matching can be measured. each document and question independently in dense vectors, and the similarity score can be calculated. the same for answer extraction neural models can also be applied. So the new architecture take a query in natural language to be answered by a QA model composed with two main components the retriever, this component is responsible for extracting relevant documents or passages from a large corpus of texts according to the input query. It uses techniques like TF-IDF Term Frequency-Inverse Document Frequency, Best Matching 25, or neural-based approaches to efficiently retrieve a subset of documents likely to contain the answer to the user's query. Once the relevant passages have been retrieved by the extractor, the reader comes into play. The reader is

tasked with understanding the retrieved text and extracting precise answers to the user's question from these passages. This often involves the use of natural language processing (NLP) techniques, such as automatic comprehension models like BERT (Bidirectional Encoder Representations from Transformers) or its variants, which are trained specifically for tasks such as answering questions by extraction.

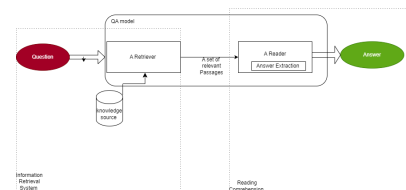


Figure 1: Architecture of QAS based on Retriever-Reader.

We propose in table 1 a new vision for building of QAS depends in the model of QA that composed from the retriever and the reader . that can give a new classification of QAS shown in 2: a open extractive book based on information retriever and and a machine reading comprehension to read answers from a given context and a open abstractive book that the answer is a text generated from a QA model.

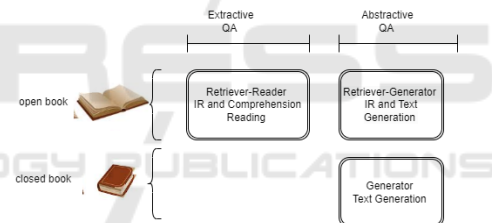


Figure 2: Approaches to Question-Answering.

## 2 RELATED WORK

Arabic Question Answering System (QAS) presents a number of challenges and peculiarities compared to a traditional QAS using other languages such as the linguistic complexity: The Arabic language has a complex grammatical and syntactic structure, with particular declension and conjugation rules. This makes automatic comprehension of questions and finding answers more difficult than in languages with simpler structures and the scarcity of linguistic resources: Unlike some Western languages, for which there are vast annotated datasets, lexicons and linguistic resources. in the last the sensitivity to cultural contexts: Questions asked in Arabic may be influenced by specific cultural and religious references. QAS must be sensitive to these contexts in order to provide appropriate answers.

Table 1: Building QAsystem.

	ExtractiveQA	Open Generative	Closed Generative
the use of Context	yes	yes	no
way of the answer is obtained	Extraction	Generation	Generation

## 2.1 Arabic Question Answering System

Question Answering (QA) tasks can be grouped into different categories according to the type of question asked: factoid, list, definition and non-factoid. Factual questions generally concern specific entities. Questions beginning with words such as When, Where, How much, What and Who are other examples of this type. In Arabic, these types of questions are often used in research. Various Arabic QAS have been proposed in the literature. QARAB is a system proposed by (Hammo et al., 2002) which takes factoid Arabic questions and attempts to provide short answers. ArabiQA proposed by Benajiba et al in (Benajiba et al., 2007) its research work is interested in modern Arabic language. It answers factoid questions based on named entity recognition. QASAL (Brini et al., 2009) is designed to answer factoid and definition questions. AQUASYS (Bekhti and Al-Harbi, 2013a) is designed to answer Arabic questions related to named entities of types person, location, organization, time, quantity. The authors assumed that the answer is a short passage. IDRAAQ (Abouenour et al., 2012) is an OpenQAS designed to enhance the quality of retrieved passages. JAWEB (Kurdi et al., 2014) is based on AQUASYS (Bekhti and Al-Harbi, 2013b). JAWEB provides a web interface to the system, which is an additional support for Arabic language presentation in web browsers. The system scored 100% recall and 80% precision. LEMAZA (Azmi and Alshenaifi, 2017) is built to answer Arabic why-questions. This system achieved about 72.7% Recall, 79.2% Precision. The SOQAL system, as described by (Mozannar et al., 2019a), is notable for being the first Arabic Open Question Answering System (OpenQAS) to employ a neural approach. It comprises two primary components: a retriever and a reader. For the retriever it uses TF-IDF technique to retrieve first a set of documents related to the question. For the reader it is a neural reading comprehension model based on BERT. SOQAL achieved 20.7%, 42.5%, and 51.7% in exact match score, f1 score, and sentence match score, respectively. Arabic QA4MRE presented by Trigui et al in (Trigui et al., 2012) introduced the Arabic language for the first time to CLEF. The system proposes a new approach that allows questions with multiple answer choices to be answered from short Arabic texts.

## 2.2 Arabic QA Datasets

In this context we present a set of Arabic Question answering datasets we propose to classify them into categories as shown in figure 2.

**ARCD:** a open book dataset Crowd-sourced composed by 1,395 questions based on 465 paragraphs from 155 articles presented by (Mozannar et al., 2019b)

**TYDIQA:** is a closed book question answering dataset covering 11 typologically diverse languages with 204K question-answer pairs presented in (Clark et al., 2020).

**Arabic-SQuAD:** a open book dataset for extractive QA system composed by 48,344 questions on 10,364 paragraphs that are translated from English SQuAD presented in (Mozannar et al., 2019b)

**MLQA:** is MultiLingual Question Answering), a benchmark dataset for evaluating cross-lingual question answering performance. It is an extractive QA proposed by (Clark et al., 2020)

**SOQAL:** is a open book dataset for extractive QA based on : a document retriever using a hierarchical TF-IDF approach and a neural reading comprehension model using the pre-trained bi-directional transformer BERT (Clark et al., 2020)

**So2al-wa-Gwab:** A New Arabic Question-Answering Dataset Trained on Answer Extraction Models presented in (Al-Omari and Duwairi, 2023)

**CQA-MD:** a medical Arabic corpus for closed book was proposed by (Nakov et al., 2019). The corpus contains over 100k questions-answers pairs collected from Arabic Medical websites.

## 2.3 Quranic Question Answering System

Many studies have building systems for locating answers to Quranic inquiries from the Holy Quran. These investigations generally employed different techniques. Abdelnasser et al. (2014) in (Abdelnasser et al., 2014) have focused in retrieval techniques to developing a open extractive book system named Al-Bayan. This system taked the question as input and then retrieves a Quran verse that includes the answer based on ontology by computing the cosine semantic similarity between the question and the concept vectors. Other researchers try to extract answers from Hadith using the same techniques, such

Table 2: Quranic QA datasets.

Dataset	Description
A corpus of Quran and tafsir (Abdelnasser et al., 2014)	A Collection of dataset from a corpus of the quran and tafsir(Ontology)
Surah al-baqarah(Ahmad et al., 2016)	Most popular English translation of the Quran of Abdullah Yusuf Ali.
Quran DB (Alqah-tani and Atwell, 2018)	Arabic-English Quran ontologies from different datasets related to Al Quran.
Ayatec (Malhas and Elsayed, 2020)	207 questions (with their corresponding 1,762 answers) covering 11 topic categories of the Holy Qur'an
QRCD (Malhas and Elsayed, 2022)	Extractive open book QA for Qur'anic Reading Comprehension Dataset
QUQA (Alnefaie et al., 2023)	2,189 questions, classified as 1778 single-answer and 411 multiple-answer questions

as the work of (Maraoui et al., 2021) with an accuracy of 92% (Hamoud and Atwell, 2016) recommend building a open extractive system on a knowledge base of related domains to answer all kinds of questions about the Qur'an. They constructed a corpus of 1,500 questions and their answers. The dataset included different types of questions. The system is based on matching the user query to the questions in the dataset in order to find the most relevant question and display its answer. This system demonstrated a precision of 79 % and a recall of 76%. The performance of this approach is affected by the size and variety of the corpus. In recent years, several studies have focused on the use of pre-trained models for the Qur'anic Machine Reading Comprehension (MRC) task. The authors used the Qur'anic Reading Comprehension Dataset (QRCD). (ElKomy and Sarhan, 2022) to build a extractive QA for open book. they tested five pre-trained models: Arabic BERT (ARBERT), AraBERTv02-Base, AraBERTv02-Large, Masked Arabic BERT (MAR-BERT) and QCRI [Qatar Computing Research Institute] Arabic and Dialectical BERT (QARiB)-Base. Their system yielded the following results: 0.27 Exact Match (EM), 0.50 F1@1 and 0.57 partial Reciprocal Rank (pRR). (Ahmed et al., 2022) augmented the QRCD with 657 questions and trained the AraElectra model. The system yielded the following results: 0.24 EM, 0.51 F1@1 and 0.55 pRR. Malhas and Elsayed (2022) in (Malhas and Elsayed, 2022) interested in reading comprehension to build an open extractive book, they pre-trained the Arabert model on data in classical Arabic and then they fine tuned it on data from the QRCD dataset. This model, called CL\_Arabert, outperformed Arabert with a partial precision of 0.51. With the advent of large lan-

guage models, the focus turned to the use of these models in NLP tasks, particularly In Question Answering system. In (Alnefaie et al., 2023) introduced the use of the GPT family to generate answers to questions from the QUQA dataset, they made just the experiments and they use a manual evaluation Performance evaluations of this system resulted in F1=0.26 for the text portion of the GPT4 generated answers and F1=0.32 for the evaluation of the Quranic series portion of the GPT-4 answers.

## 2.4 Quranic QA Datasets

We present a summary of Qur'anic datasets in the table 2.

## 2.5 Large Language Models

Large Language Models (LLMs) such as the GPT series mentioned by research (Radford et al., 2018). GPT's progression has already gone through several generations: GPT-1 (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and GPT-4 (OpenAI, 2023). They have been at the forefront of recent NLP research following their ability to generate consistent and contextually accurate textual results. These models have demonstrated impressive performance in natural language processing (NLP) tasks ranging from summarization and translation to question answering. The strategy we are going to implement in this article to question LLM is the prompt engineering in order to guide our Quranic question answering system to produce good results.

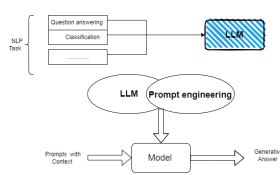


Figure 3: LLM for QA.

## 2.6 Prompt Engineering

Prompt engineering is the practice of designing efficient inputs, or prompts, to achieve desired results from AI language models such as LLMs. It has many objectives such as:

- **Improved performance:** Well-designed prompts can help AI models produce more accurate, relevant and consistent results. This is particularly important in applications where precise, nuanced or contextual responses are required.
- **User experience:** Effective prompt engineering contributes to a better user experience by enabling AI models to understand the user’s intent more precisely, and to respond in a more useful and compliant way to their expectations.
- **Risk mitigation:** Carefully designed prompts can reduce the likelihood of AI models generating inappropriate content by guiding them towards safer, more context-aware results.
- **Domain adaptation:** Prompt engineering enables AI linguistic models to be adapted more effectively to specific domains or applications, even when they have been trained beforehand on general linguistic data ,in our case, this is Quranic question-answer systems.

### 2.6.1 Techniques of prompt Engineering

there is a specific types of prompting depends in context of application of this technique.

- **Zero\_shot Prompting:** With this technique we do not provide the LLM with examples of texts in parallel to get better answers.
- **Few-shot Prompting:** is a technique where the input consists of a natural language query or instruction, followed by one or more examples of the desired output.
- **Chain of Thought:** Wei et al (2022) proposed the "chain of thought" (CoT) prompt, which achieves complex reasoning abilities through intermediate reasoning steps.
- **Self.Consistency:** Introduced by Wang et al (2022), self-consistency aims to "replace naive

decoding in chain-of-thought prompts".It improves the performance of chain-of-thought prompting in tasks involving arithmetic reasoning.

## 3 PROPOSED APPROACH

Our proposed approach consists in building a Open-book abstractive Quranic question answering system. we will build a reader-retriever model and explore its capabilities in providing accurate and well-sourced answers to our Quranic questions based on the technique of zero\_shot prompting using GPT-4-turbo LLM tested the QUQA dataset. Our approach is to ameliorate the output of the retriever and restrict the relevant passages to a verses of the Qur’an and generate answers in a specific format where it will be easy to apply automatic evaluation. The QUQA contains three kinds of question. The QUQA contains three types of questions: a 103 confirmation questions that the answer can be yes or no those question begin with هل , a 1,621 descriptive questions started with

ما ، كم ، كيف ، لماذا ، ماذا ، أذكر

their answers can be reason explanation or a definition and 465 factoid questions begin with

من ، أين ، متى

with a answer type a location, a duration or a name of person . The version used for this study was GPT-4-turbo, it is latest version of the GPT at the time of our research. This model is used directly and does not require any fine-tuning process but it needs guidance We used the Python API to retrieve answers from the model. Google Colab was used to run our experiments.

### 3.1 Metrics of Evaluation

The output generated by GPT-4-turbo ,our LLM in response to our provided questions produces natural-sounding text, often including several series of Quranic verses, These series may vary in length, comprising one or more verses. Consequently, we conducted separate evaluations for the textual answers and the Quranic verses themselves.The evaluation of the Quranic verses involved both automated processes.by prompting the system to display a ranked list of the Quranic verses in the answer, we extract a list of predictions and compare them with different references answers mentioned as gold responses. In order to evaluate the generated responses, we used the evaluation metrics mentioned in the research work

of (Malhas and Elsayed, 2022), we used F1 which is a measure of a test's accuracy that considers both the precision and recall of the test to compute the score. To compute it, we measured the F1 between each series of verses and the golden answer or the list of golden answer for the question multi answers and then took the average. The Exact matching EM and the F1@1 are calculated in relation to the answer at rank 1. For F1@1 we take the maximum if the question has many golden answers.

## 4 RESULTS

In this section we present the results of our experiments, we try several prompts in the query of our model, the different prompts generate different responses. We carried out three large-scale experiments on the QUQA dataset, dividing the experiments by question type. Each large-scale experiment was repeated several times, each time modifying the prompt system. In the first experimentation we set just the user prompt regardless of the question asked, then in the remaining experimentation we defined a system prompt to improve the evaluation results. The results of the first experiment are shown in the table 3, while the results of the second experiment, in which the prompt system was mentioned.

Example 1 of prompt of system :

اجب عن السؤال حسب المراجع الاسلامية من القرآن الكريم

are shown in table 4.

The results of evaluation of Example 2 of prompt of system:

اجب على السؤال المطروح بذكر قائمة أدلة من القرآن الكريم

are shown in table 5

Table 3: Results of first experimentation prompt of user = Question.

	Fscore	EM	F@1
Confirmation	0,17	0,11	0,19
Factoid	0,25	0,15	0,23
Descriptive	0,31	0,19	0,25
All	0,24	0,15	0,22

Table 4: Results of prompt 1 of System.

	Fscore	EM	F@1
Confirmation	0,22	0,13	0,21
Factoid	0,25	0,23	0,32
Descriptive	0,33	0,20	0,28
All	0,26	0,18	0,27

Table 5: Results of evaluation of our prompt2 of System.

	Fscore	EM	F@1
Confirmation	0,34	0,20	0,30
Factoid	0,38	0,25	0,32
Descriptive	0,40	0,28	0,35
Our evaluation of (Alnefaie et al., 2023)	<b>0,37</b>	<b>0,24</b>	<b>0,32</b>
	0,32	0,19	0,26

## 5 ANALYSE AND DISCUSSION

The QUQA dataset contains three kinds of question: a 103 confirmation questions that the answer can be yes or no those questions begin with هل , a 1,621 descriptive questions started with

ما ، كم ، كيف ، لماذا ، ماذا ، أذكر

their answers can be reason explanation or a definition and 465 factoid questions begin with

من ، أين ، متى

with a answer type a location, a duration or a name of person.

This section analyses and discusses the results of the evaluation of our system based on the prompting technique. In first case ,the prompt contains just the question without any context In the Second case, the results obtained for the confirmation questions are improved by the quality of the prompting. In this case, it is important to guide the model to find the Qur'anic verses corresponding to the answer. In the third case ,Our prompt system 2 is a conditional text generation system based on prompts. This technique enables the generation of controlled and contextually relevant text, making it useful for the evaluation system.

## 6 CONCLUSION

In this paper, we have evaluated a Quran question answering system based on LLM specially the GPT4-turbo version ,our main objective is to know how the evaluation results as evolve as a function of prompting. As a future work, we aim a generalization to Continuous our experimentation using other techniques of prompting like few shot or chain of taught combined

with other techniques like Retrieval Augmented Generation (RAG) for LLMs on other datasets.

## REFERENCES

- Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N. M., and Torki, M. (2014). Al-bayan: an arabic question answering system for the holy quran. pages 57–64.
- Abouenour, L., Bouzoubaa, K., and Rosso, P. (2012). Idraaq: New arabic question answering system based on query expansion and passage retrieval.
- Ahmad, N. D., Bennett, B., and Atwell, E. (2016). Semantic-based ontology for malay qur'an reader.
- Ahmed, B., Saad, M., and Refaee, E. A. (2022). Qqateam at qur'an qa 2022: Fine-tuning arabic qa models for qur'an qa task. pages 130–135.
- Al-Omari, H. and Duwairi, R. (2023). So2al-wa-gwab: A new arabic question-answering dataset trained on answer extraction models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–21.
- Alnefaie, S., Atwell, E., and Alsalka, M. A. (2023). Is gpt-4 a good islamic expert for answering quran questions? pages 124–133.
- Alqahtani, M. and Atwell, E. (2018). Annotated corpus of arabic al-quran question and answer.
- Azmi, A. M. and Alshenaifi, N. A. (2017). Lemaza: An arabic why-question answering system. *Natural Language Engineering*, 23(6):877–903.
- Bekhti, S. and Al-Harbi, M. (2013a). Aquasys: A question-answering system for arabic. 25(6):19–27.
- Bekhti, S. and Al-Harbi, M. (2013b). Aquasys: A question-answering system for arabic. 25(6):19–27.
- Benajiba, Y., Rosso, P., and Lyhyaoui, A. (2007). Implementation of the arabiqa question answering system's components. pages 3–5.
- Brini, W., Ellouze, M., Mesfar, S., and Belguith, L. H. (2009). An arabic question-answering system for factoid questions. pages 1–7.
- 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.
- Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- ElKomy, M. and 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*.
- Hammo, B., Abu-Salem, H., Lytinen, S. L., and Evens, M. (2002). Qarab: A: Question answering system to support the arabic language.
- Hamoud, B. and Atwell, E. (2016). Quran question and answer corpus for data mining with weka. pages 211–216.
- Kurdi, H., Alkhaider, S., and Alfaifi, N. (2014). Development and evaluation of a web based question answering system for arabic language. *Computer science & information technology (CS & IT)*, 4(02):187–202.
- Malhas, R. and 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(6):1–21.
- Malhas, R. and Elsayed, T. (2022). Arabic machine reading comprehension on the holy qur'an using cl-arabert. *Information Processing & Management*, 59(6):103068.
- Maraoui, H., Haddar, K., and Romary, L. (2021). Arabic factoid question-answering system for islamic sciences using normalized corpora. *Procedia Computer Science*, 192:69–79.
- Mozannar, H., Hajal, K. E., Maamary, E., and Hajj, H. (2019a). Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.
- Mozannar, H., Hajal, K. E., Maamary, E., and Hajj, H. (2019b). Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.
- Nakov, P., Márquez, L., Moschitti, A., and Mubarak, H. (2019). Arabic community question answering. *Natural Language Engineering*, 25(1):5–41.
- OpenAI, R. (2023). Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Trigui, O., Hadrich Belguith, L., Rosso, P., Ben Amor, H., and Gafsaoui, B. (2012). Arabic qa4mre at clef 2012: Arabic question answering for machine reading evaluation.