AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text

Yousef Adel, Mostafa Dorrah, Ahmed Ashraf, Abdallah ElSaadany, Mahmoud Mohamed,
Mariam Wael, Ghada Khoriba
Center for Informatics Science (CIS), School of Information Technology and Computer Science (ITCS),
Nile University, Giza, Egypt
{Y.Khalil, M.Samer, Ahme.Ashraf, A.Elsaadany, mahmoud.mohamed, Mar.Wael, Ghada Khoriba}@nu.edu.eg

***Abstract*—Recently, the internet has become a vast repository of religious texts and sources. The quest for valid and dependable Islamic Q/A that provides accurate substantiation from the Holy Quran (Muslim holy book) and Hadith (Prophet Muhammed teachings) has become challenging, given the abundance of misleading answers lacking credible evidence and proper sources. Concurrently, transformer-based architectures have demonstrated remarkable efficacy in language modeling and comprehension. Notably, applications in religious Arabic generative question answering have remained underdeveloped, primarily due to the lack of available Arabic religious datasets. In this paper, we present an Arabic Islamic generative question-answer model named AraQA, which has been fine-tuned using Arabic Islamic question/answer pairs meticulously gathered and extracted from reputable open-source websites on the internet. The model is initially designed to operate exclusively in the Arabic Language. Our model attains an impressive perplexity score of 2.3 when evaluated on held-out question-answer pairs. We have made the model publicly accessible via GitHub, anticipating it will pave the way for research in Arabic and religious Natural Language Processing (NLP).** [1]

***Index Terms*—Transformers, Question-Answering, Religious Text, Arabic NLP, Topic Modeling, Large Language Models, GPT**

I. Introduction

In recent years, the field of NLP has witnessed a groundbreaking advancement with the emergence of the multi-headed self-attention transformer architecture, famously known as "attention is all you need" [1]. This transformer-based approach has propelled NLP to new heights, outperforming traditional Recurrent Neural Networks (RNNs) in various tasks. From sentence completion using GPT2 [2] to sentence classification and machine translation with BERT [3], the transformer has demonstrated its effectiveness in revolutionizing language understanding. Notably, the success of these models has been further amplified by leveraging larger architectures and training on vast datasets. The likes of RoBERTa [4] and XLM-R [5] have shown that scaling up models and data can yield even more impressive performance gains. Among this remarkable progress in NLP, addressing the pressing needs of diverse Arabic communities and cultures is essential. Arabic, with over 313 million speakers and a profound connection to over 1.9 billion Muslims worldwide, represents a significant linguistic and religious context. There remains a need for comprehensive guidance in Islamic Sharia, particularly in the provisions related to the pillars of Islam, such as prayer, fasting, and more. In today's digital age, the internet is filled with misleading answers. Searching for authenticated information based on the Quran and the Hadith of Prophet Mohamed is difficult. Thus, we proposed our model (AraQA: [**?**]), built upon the foundation of the cutting-edge AraGPT-2 architecture [6]. The proposed GPT-based model has been trained on various Islamic documents, including the Qur'an, Hadith, and Books of Islamic law (fiqh) written by some of the best Muslim jurists throughout history. In addition, our collected dataset contains many frequently asked queries and their authenticated responses. The proposed model's broad knowledge base allows it to understand and respond to various questions about Islam while providing accurate and contextually relevant answers. The proposed model's extensive knowledge base enables it to comprehend and reply to various inquiries regarding Islam while giving contextually appropriate responses with a perplexity score of 2.3. The perplexity score was evaluated on held-out question-answer pairs. As a result of our proposed model, we propose an Islamic generative chatbot for answering all Islamic questions on all the topics in the main Islamic Law books.

This paper is organized as follows: Section 2 discusses Arabic and Islamic Q/A systems-related works. Section 3 presents the proposed methodology for constructing our Q/A dataset and model architecture. Section 4 provides our experimental setup, results, and discussion. Section 5 briefly discusses our previous and current experiments' results. In Section 6, we conclude the findings of our experiments, limitations, and future work.

II. Related Work

Arabic Question and Answering (QA) systems have witnessed notable advancements in recent years, driven by the integration of machine learning, natural language processing, and information retrieval techniques. Several related systems have emerged within this domain, each aiming to address unique challenges and requirements.

Mohammed et al. [7] introduced the Arabic Question Answering System (AQAS) as one of the pioneering Arabic question-answering systems. This system processes Arabic sentences, encompassing declarative statements and questions,

[1] https://github.com/Marje3na/AraQA-An-Arabic-Generative-Question-Answering-Model-for-Authentic-Religious-Text

to generate desired user responses. It leverages the frames technique to represent the knowledge base of a specific domain, contributing to the broader landscape of NLP systems. Hamed et al. [8] Proposed a novel approach to develop an efficient Question Answering System (QAS) tailored to retrieve accurate Holy Quran verses. This system utilizes the Neural Network (NN) technique and a dataset comprising the popular English translation of the Quran by Abdullah Yusuf Ali [9]. It incorporates question expansion using WordNet and integrates Islamic terms to address translation discrepancies [10]. The QAS classifies Quranic surahs, such as Al-Baqarah, into categories, streamlining the retrieval of relevant verses. Moreover, the system ranks these verses based on similarity scores, yielding high accuracy. The evaluation of the NN classification has shown a 90% level of accuracy, while the proposed approach based on the entire QAS has an 87% level of success.

Altammami and Atwell [11] explored the application of Transformer-based models to the intricate realm of classical Arabic manuscripts, specifically the Quran and Hadith. These texts demand deep human comprehension due to their complex and nuanced nature. The research evaluated state-of-the-art monolingual, bilingual, and multilingual models to establish relatedness between Quranic verses and Hadith teachings, thereby addressing a significant gap in the use of advanced models for traditional Arabic texts. Arabic systems have been proposed as QARAB [12] is a QA system that takes Arabic questions and attempts to provide short answers. ArabiQA [13], which is fully oriented to the modern Arabic language. It also answers factoid questions using Named Entity Recognition. However, this system is not completed yet. This system proposed a new approach that can answer questions with multiple answer choices from short Arabic texts. However, its overall accuracy is 0.19. To the best of our knowledge, no previous research was proposed for the Quranic classical Arabic question answering or Quran-extracted Q/A.

Alkhurayyif and Sait [14] proposed a state-of-the-art model in open-domain Arabic question/answering, with 96.23% accuracy on the ARCD dataset and 95.35% accuracy on the TyDiQA dataset. Their proposed model used three approaches: multinomial naïve Bayes algorithm to classify the content of the datasets used, such as the topic, question, and text. Then, the model used the embedding from the language model (ELMo) to convert the user query into vectors of numbers to identify the existence of the current user query in the query storage module to respond in a limited time. ELMo outperforms other embedding techniques, such as GloVe and Word2Vec. The ELMo uses QLSTM to capture multidimensional features, thus more accurate embeddings within the context. However, the limitation of their proposed model is that it needs additional training to cover more different Arabic terms as the verb + noun term patterns, in addition to the NER, need to be improved by fine-tuning to capture language ambiguities. Lastly, the proposed model, with the entirety of its components, cannot be further used and extended for any other NLP applications.

In light of these significant contributions to the Arabic QA landscape, a compelling need remains for specialized systems that address specific domains, such as Islamic knowledge. The model proposed in this paper, "AraQA," is initially trained on our collected datasets. While the model's primary language is Arabic, it has the potential to enhance accessibility to Islamic knowledge, offering Muslims accurate and relevant answers to their inquiries while reducing reliance on individual scholars or clerics. This new method could potentially improve the way Muslims access Islamic information, contributing to a community that is more knowledgeable and engaged in their quest for understanding. AraQA is a generative Arabic chatbot that can answer any Islamic query. AraQA is based on the modern AraGPT-2 architecture and has been trained on various Islamic resources, including the Qur'an, Hadith, Books of Islamic law (fiqh), and a dataset containing multiple commonly asked questions and their answers. The depth of AraQA's knowledge base enables it to answer a wide range of topics about Islam and respond with accurate and, depending on context, appropriate responses. AraQA is a significant improvement over previous Arabic QA systems. It is more accurate, comprehensive, and user-friendly.

III. Methodology

Dataset, Pre-processing, Topic Modelling, Model, and Validation of Aya and hadith are the five main components of our proposed solution. In the training time, as shown in Figure 1, the scrapped data was pre-processed, the questions were categorized according to their topics, and then used to fine-tune the AraGPT2-Base model. In the Inference time, as shown in Figure 2 the query was categorized using topic modeling and then passed to the trained model to generate the answer. Then, the generated answer is passed through a validation method that validates that the verse and hadith in the generated answer are not structured destroyed.

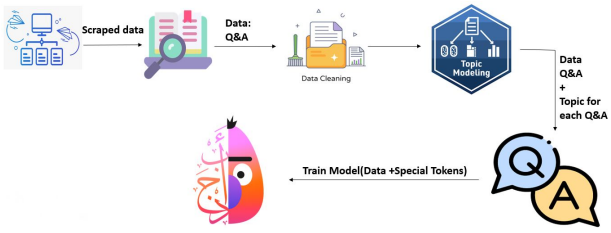

Fig. 1. Model Architecture

A. Dataset

The training dataset is a question/answer pair collected and scraped from a trusted website on the internet. 88,606 pairs were collected in total. The topics that the questions in the data (after some data cleaning) belong to are shown in Figure 3.

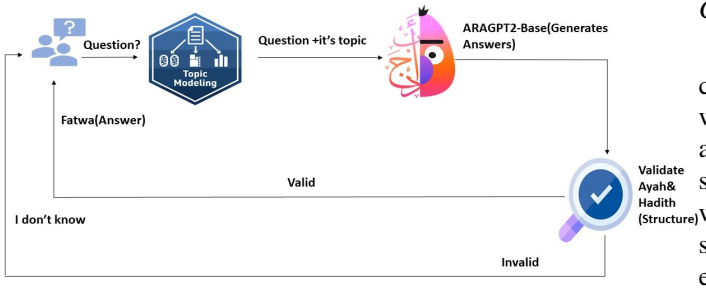

Fig. 2. Inference Time

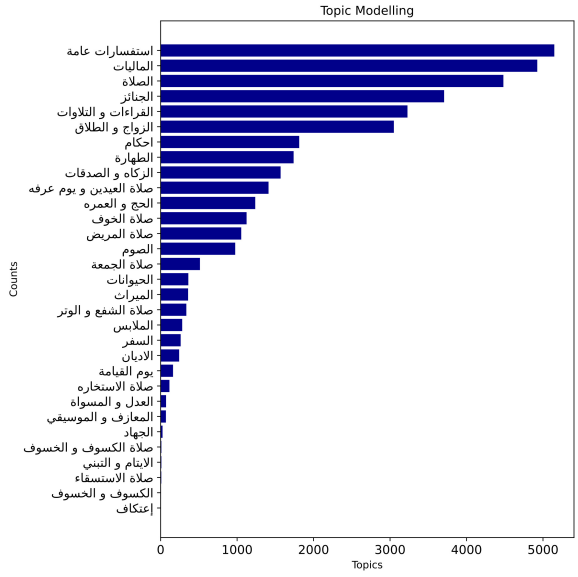

Fig. 3. Most frequent topics

B. Pre-Processing

The data is filtered by removing NULL values, irrelevant text that does not represent the data completely, and text containing inaccessible references using a function that detects phrases such as "فتوى رقم" "انظر الى الفتوى", and short text with less than 118 characters are also filtered. It was found that 118 is the optimal value by going through the data manually where all of the answers with a size less than 118 contain only the opening phrase of the answer which does not hold any contextual meaning to the question asked. Ambiguous, unnecessary characters are replaced with an empty string using a function that extracts only the Arabic characters with English digits, punctuations, and brackets, while any other character does not pass the filtering function. Hence, the text is cleaned. Moreover, the question is concatenated with the answer in one string and adding a special token to the tokenizer at the beginning of each question and the beginning of each answer. After filtering and preprocessing, the new size of the data is 40,859 pairs of questions/answers.

C. Topic Modelling

Given the specific characteristics of our dataset, which consists of short textual content related to the Islamic domain, we decided to use regular expressions for topic modeling to achieve the best results. Regular expressions are powerful string-matching patterns that allow us to identify specific word patterns or phrases within the text. Our topic modeling strategy depends on identifying the main Islamic topics. For each topic category, we compiled a set of keywords that are associated with that topic. These keywords were selected based on our domain knowledge and the most frequent keywords (dominant words indicative of particular topics). We segmented our analysis into sub-topics and main topics to capture variations within specific topics. For example, within the broader topic of الصلاة. we identified sub-topics such as صلاة الاستسقاء ، صلاة الاستخارة ، صلاة الخوف and more. Applying this approach, we could categorize short textual content into distinct topics. As shown in Figure 4, each question has multiple keywords, and according to the conditions of our topic modeling algorithm, each keyword in the questions has a degree of dominance, while a darker color is more dominant than a brighter color. As shown in the first example, the keyword "هل يجوز" is under the topic of احكام, and "الصلاة" is under the topic of الصلاة. However, the topic modeling algorithm considers the question to be categorized as a question in the topic of الصلاة. However, الصلاة is the dominant keyword in the question despite being at the end of the sentence. So, our topic modeling algorithm does not consider the keyword's position.

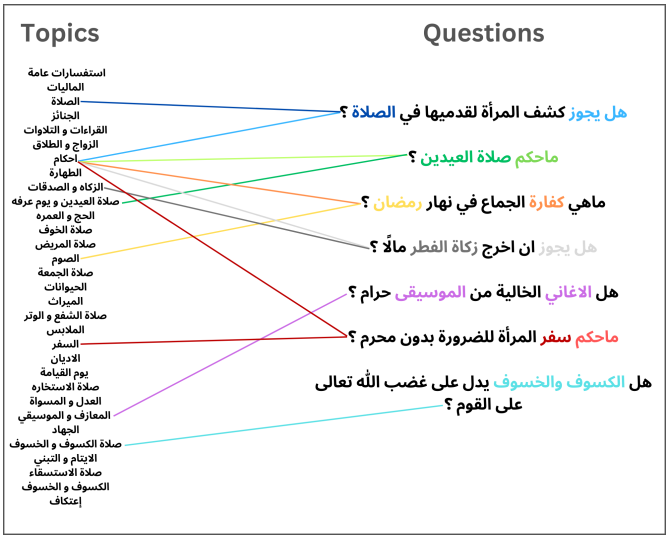

Fig. 4. Topic modeling

D. Proposed Model

ARAGPT2-Base version is used to fine-tune the proposed model. AraQA is trained with a context size of 1024. The

optimizer used for the model is the LAMB optimizer since it allows large batch sizes without worrying about training divergence [6]. The model has an embedding size of 768, with 12 layers and 12 heads. ARAGPT2-Base has a size of 135M parameters.

E. Validation

As shown in Figure 5, the validation process of the generated Hadith and verses involves embedding both the authentic verses and Hadith dataset and the generated verses and Hadith using the "E5" Base model [15]. Subsequently, cosine similarity between the embeddings was calculated, allowing for a quantitative measure of their structural similarity. A predefined threshold for cosine similarity was established; if the cosine similarity exceeded 90 percent, the generated verses and Hadith were replaced by the most similar verses and Hadith from the dataset. However, if the threshold was not met, the data was considered dissimilar and thus excluded from further consideration.

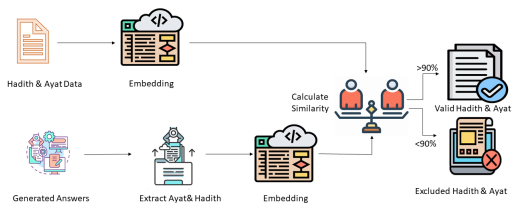

Fig. 5. Validation of Generated verses and Hadith

IV. Experiments and Results

The experiments for the proposed model are done on a single GPU (NVIDIA GeForce RTX 3070 8GB GDDR6) with a total number of steps of 136,100 in 19 epochs, with a training batch size of 10 and a learning rate of 5e-05. Due to a lack of resources and computational power, the training process of the proposed model was conducted using a random subset of 27,778 pairs of question and answer, 1,462 pairs for the validation process, and 100 held-out unseen pairs for the testing process. The proposed model was trained for 19 epochs, approximately seven hours and four minutes only, due to hardware access limitations. The cross-entropy loss function is used for the training/validation loss. The proposed model achieved a training loss after 19 epochs of 2.238 and a validation loss after 19 epochs of 3.030. The perplexity score was selected as an evaluation metric for the testing process since it calculates the degree of 'uncertainty' a model has, assigning probabilities to the test pairs of questions and answers. The proposed model achieved a mean perplexity score of 5.221. The following figure shows a sample of the proposed model's answers to specific questions.

Generated Answer	Prompt
الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد: فإن التعامل بالربا محرم بالكتاب والسنة والإجماع، قال تعالى (يا أيها الذين آمنوا اتقوا الله وذروا ما بقي من الربا إن كنتم مؤمنين * فإن لم تفعلوا فأذنوا بحرب من الله ورسوله وإن تبتم فلكم رؤوس أموالكم لا تظلمون ولا تظلمون) [البقرة: 278 - 279]. وقال صلى الله عليه وسلم: "لعن الله آكل الربا وموكله وكاتبه وشاهديه وقال: هم سواء" رواه مسلم وأحمد عن جابر بن عبد الله رضي الله عنهما. والله أعلم.	ماحكم الربا
الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد: فلا يجوز للمسلم أن يشرب الخمر، لقوله تعالى: (يا أيها الذين آمنوا إنما الخمر والميسر والأنصاب والأزلام رجس من عمل الشيطان فاجتنبوه لعلكم تفلحون) [المائدة: 90]. ولقوله صلى الله عليه وسلم: "لعن الله الخمر وشاربها وساقيها وبائعها ومبتاعها وعاصرها ومعتصرها وحاملها والمحمولة إليه وآكل ثمنها. رواه أبو داود والترمذي وقال حديث حسن صحيح. والله أعلم.	ماحكم شرب الخمر

Fig. 6. AraQA's Sample Output

V. Discussion

A previous experiment was done without the use of topic modeling, additional special tokens, and without additional data cleaning. It has been found that the use of the special tokens is very effective due to a high decrease in the training/validation loss. The model's first training loss was 4.9, while the model started with a training loss of 3.2 when special tokens were provided and is learning much faster. The model finished the training/validation process with a training loss of 1.9 and a validation loss of 2.6 in 30 epochs. According to the results shown in the previous section and the results in the previous experiment, in our final experiment, the minimum validation loss the model achieved was 2.959 in the 8th epoch, then started to increase. In our previous experiment, the validation loss in the 8th epoch was 2.560, which is lower than the current experiment. Thus, in our current experiment, with the topic modeling, the additional special tokens, and the additional data cleaning, the dataset becomes more complex for the model and smaller due to the additional data cleaning and the computational hardware limitations that limit us to using a larger model. In our previous experiment, the perplexity score was 2.3, while in the current experiment, the perplexity score on 100 held-out-set is 5.221, which is higher and worse, which justifies our argument in this section.

VI. Conclusion and Future Work

We have proposed AraQA, the first generative model to approach religious text generation, especially Islamic text. We collected data from different sources and applied several pre-processing approaches to prepare the data for fine-tuning the model. We used the ARAGPT2-Base model to fine-tune it on our collected data. For the evaluation, cross-entropy loss function and perplexity measure are used to evaluate the model's performance. In future work, we will use a larger model with more parameters to capture more complex features in the data, in addition to seeking help from a domain expert to

improve our dataset and validate the model's performance for human evaluation. In our proposed model, "AraQA", suffers from multiple limitations and can be improved in the future. One of the limitations is handling long questions and the lack of an Islamic corpus that is useful for our objective. The inability to reach a domain expert for human evaluation is a significant limitation. Moreover, the inference time is approximately 20 seconds on average, which is considered a long waiting time.

Ethics Statement

There is no ethical conflict in our research.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: a robustly optimized bert pretraining approach (2019)," *arXiv preprint arXiv:1907.11692*, vol. 364, 1907.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[6] W. Antoun, F. Baly, and H. Hajj, "Aragpt2: Pre-trained transformer for arabic language generation," *arXiv preprint arXiv:2012.15520*, 2020.

[7] F. Mohammed, K. Nasser, and H. M. Harb, "A knowledge based arabic question answering system (aqas)," *ACM SIGART Bulletin*, vol. 4, no. 4, pp. 21–30, 1993.

[8] S. K. Hamed and M. J. Ab Aziz, "A question answering system on holy quran translation based on question expansion technique and neural network classification.," *J. Comput. Sci.*, vol. 12, no. 3, pp. 169–177, 2016.

[9] A. Y. Ali *et al.*, *The Holy Qur'ān: text, translation and commentary*. Islamic Foundation, 1975.

[10] T. Hao, W. Xie, and F. Xu, "A wordnet expansion-based approach for question targets identification and classification," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, Guangzhou, China, November 13-14, 2015, Proceedings 14*, pp. 333–344, Springer, 2015.

[11] S. Altammami and E. Atwell, "Challenging the transformer-based models with a classical arabic dataset: Quran and hadith," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1462–1471, 2022.

[12] B. Hammo, H. Abu-Salem, S. L. Lytinen, and M. Evens, "Qarab: A: Question answering system to support the arabic language," in *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, 2002.

[13] Y. Benajiba, P. Rosso, and A. Lyhyaoui, "Implementation of the arabiqa question answering system's components," in *Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April*, pp. 3–5, 2007.

[14] Y. Alkhurayyif and A. R. W. Sait, "Developing an open domain arabic question answering system using a deep learning technique," *IEEE Access*, 2023.

[15] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," 12 2022.