Leveraging Language Models for Islamic Question and Answering System

Lama A. AlQarni
KAUST
lalqarni0014.stu@uj.edu.sa

Lina M. Alqahtani
KAUST
linaalqahtaniii@gmail.com

Haifaa M. Althobaiti
KAUST
haifaalthobaiti15@gmail.com

Asmaa A. Mahdi
KAUST
asmaa.qhtaniii@gmail.com

Reema F. Aljalajel
KAUST
reemaaljalajel@gmail.com

Supervised By:
Muhammad Mubashar
19230664@brookes.ac.uk

***Abstract*—In recent years, advancements in Natural Language Processing (NLP) have been significantly driven by large language models (LLMs) capable of generating and understanding human-like text. This research explores the application of LLMs in the domain of Islamic question answering, a field characterized by its complexity and sensitivity due to the intricacies of religious texts such as the Quran and Hadith. The study investigates the performance of two state-of-the-art pre-trained models, AraT5 and AraBART, in addressing questions related to Islamic jurisprudence and theology, and compares their effectiveness with a baseline Transformer model, TransArab. AraT5, tailored for Arabic text, and AraBART, known for its strong text generation and comprehension abilities, were fine-tuned on a dataset of 83,000 Islamic fatwas sourced from Islamweb. Evaluation metrics were employed to assess the models' performance, including Precision, Recall, F1-Score, and BLEU Score. Results indicate that fine-tuning significantly enhances model effectiveness, with Fine-Tuned AraBART achieving the highest Precision, Recall, F1-Score, and BLEU scores. The research underscores the importance of fine-tuning for domain-specific applications and highlights the potential and limitations of LLMs in handling complex, culturally sensitive questions. Future work will focus on optimizing model performance by experimenting with various LLM sizes and expanding the dataset to cover broader aspects of Arabic question answering.**

I. Introduction

In recent years, the field of Natural Language Processing has witnessed significant advancements, particularly with the development of large language models capable of understanding and generating human-like text. These models have shown remarkable potential in a wide range of applications, including question answering systems, which are increasingly being employed in various domains to provide accurate and contextually relevant information. Among these domains, the realm of Islamic question answering presents unique challenges and opportunities due to the richness, complexity, and sensitivity of the religious texts and principles involved.

Islamic question-answering systems require a deep understanding of the Islamic scriptures, including the Quran, Hadith, and various scholarly interpretations. Traditional methods of answering religious queries often involve consultation with scholars or extensive research into religious texts. However, with the advent of LLMs, there is an opportunity to develop automated systems that can assist in answering questions related to Islamic jurisprudence, theology, and ethics, while maintaining the necessary rigor and respect for the subject matter.

This research aims to explore the capabilities of large language models in the context of Islamic question answering by leveraging two state-of-the-art pre-trained models AraT5 and AraBART. AraT5, a variant of the T5 model specifically designed for Arabic text, and BART, a model known for its strong performance in text generation and comprehension, will be fine-tuned on Islamic texts to assess their effectiveness in providing accurate and contextually appropriate answers to Islamic questions. In addition to these fine-tuned models, a baseline transformer model will be developed to serve as a point of comparison, enabling a thorough evaluation of the performance of these advanced models in this specialized domain.

The primary objectives of this research are to assess the performance of these models in generating high-quality answers to Islamic questions, to compare their results against a baseline transformer model, and to identify any limitations or ethical considerations that may arise when deploying such models in sensitive religious contexts. Through this work, we seek to contribute to the growing body of research on the application of LLMs in religious and culturally significant domains, providing insights into the potential and challenges of using AI in Islamic question answering systems.

II. Related Work

There are numerous published studies that have advanced Arabic NLP in areas like question and answering and text generation.

Nagoudi et al. [1] developed three Arabic-specific variants of the T5 model, namely AraT5MSA, AraT5TW, and AraT5, to address the challenges in Arabic language generation and understanding, particularly focusing on tasks like question answering. The authors used a comprehensive dataset comprising 70GB of Modern Standard Arabic (MSA) text from

sources such as AraNews, Gigaword, and Wikipedia Arabic, along with 178GB of Arabic tweets that included a wide variety of dialects and naturally occurring code-switching between Arabic and other languages. The models were trained using the T5Base architecture with a self-supervised denoising objective, where the models reconstructed original sentences from corrupted versions. These AraT5 models significantly outperformed the multilingual mT5 model across various Arabic language tasks and set new state-of-the-art benchmarks on the ARGEN evaluation suite, which covers tasks like machine translation, text summarization, and question generation, as well as on the ARLUE benchmark for Arabic language understanding, showcasing their superiority in both language generation and understanding.

Ateeq et al. [2] developed mBART and mT5 models to address the gap in Arabic narrative question-answering. The dataset included the newly created Arabic-NarrativeQA dataset, which combined translated and original Arabic stories. The problem was the lack of advanced Arabic narrative comprehension models, particularly for generating accurate, contextually appropriate answers to questions based on Arabic stories. The approach involved fine-tuning mBART and mT5 models on an English NarrativeQA dataset, followed by further fine-tuning on the Arabic-NarrativeQA dataset to enable cross-lingual transfer learning. The results showed that mBART, especially after cross-lingual training, significantly improved performance, achieving a ROUGE-L score of 0.34 on the test set, effectively generating human-like answers to narrative questions. The models and dataset are publicly available, promoting further research in Arabic narrative QA.

Antoun et al. [3] developed ARAGPT2 and AraBERT models to address the gap in Arabic question-answering and language generation. The dataset included a large-scale Arabic corpus combining sources like the OSCAR corpus, Arabic Wikipedia, and news articles. The problem was the lack of advanced Arabic language generation models, particularly for question-answering tasks. The approach involved pre-training the ARAGPT2, a GPT-2-based model, from scratch in four sizes, and using AraBERT, another transformer-based model, for comparison. The results showed that ARAGPT2, especially the mega variant, successfully generated human-like text, achieving a perplexity of 29.8 on Wikipedia and correctly answering 25% of trivia questions in a zero-shot setting. The models are publicly available, promoting further research in Arabic NLP.

The paper written by Abdul-Mageed et al. [4] presents ARBERT and MARBERT, two Arabic-specific transformer-based language models, designed to handle the diversity of the Arabic language, including its dialects and social media variations. Alongside these models, the authors introduce ARLUE, a benchmark comprising 42 datasets for evaluating Arabic language understanding. Both models outperform existing multilingual models like mBERT and XLM-RLarge, achieving state-of-the-art results in 37 out of 48 tasks across six clusters, including sentiment analysis, named entity recognition, and question answering. MARBERT-v2, an enhanced version with longer sequence pre-training, significantly improved performance in question answering tasks, achieving the highest ARLUE score of 7740, showcasing the models' superior effectiveness and efficiency in Arabic NLP tasks.

III. Problem Statement

Arabic Question Answering (QA) systems have garnered significant interest due to the growing demand for accurate and reliable information across various domains, particularly within Arabic-speaking communities. The Arabic language, with its complex linguistic structure, rich morphology, syntax, and extensive vocabulary, presents unique challenges for NLP. Arabic is a Semitic language characterized by root-based morphology, where words are often formed through the combination of roots and patterns, leading to a vast array of word forms. Additionally, the language's script, diacritical marks, and the presence of numerous dialects further complicate the development of NLP systems capable of understanding and processing Arabic text accurately.

Given these challenges, the development of advanced QA systems that can effectively interpret and respond to queries in Arabic is crucial. Traditional NLP models have often struggled with these linguistic complexities, leading to the need for more sophisticated approaches. The rise of large language models in NLP offers a promising solution. These models, which include AraT5 and BART have been pre-trained on extensive Arabic corpora, enabling them to capture the nuances of the language more effectively than previous approaches.

Additionally, our QA system is designed to address complex Islamic issues by accurately interpreting religious texts like the Quran and Hadith, ensuring culturally and religiously appropriate responses to queries.

IV. Methodology

We will be following the methodology outlined in Figure 2, which involves a structured approach to processing and evaluating different models for Arabic language tasks. The process begins with Data Preprocessing, serving as the foundation for all subsequent steps by preparing the data for model input. After preprocessing, the data is distributed to three distinct models: AraT5, AraBART, and a custom Transformer model named TransArab. AraT5 and AraBART are fine-tuned by training them on the assigned dataset, while TransArab, developed following the "Attention is All You Need" paper [5], is trained from scratch. Each model undergoes a two-step process: Model Training, where the models are trained on the processed data, and Model Evaluation, where the performance of each trained model is assessed. This approach ensures a comprehensive evaluation, allowing for a comparison of each model's effectiveness in handling specific Arabic language tasks.

A. Data Acquisition and Data Preprocessing

The dataset utilized in our research was sourced from Kaggle, where it was previously scraped from Islamweb (islamweb.net). This comprehensive dataset consists of 83,000

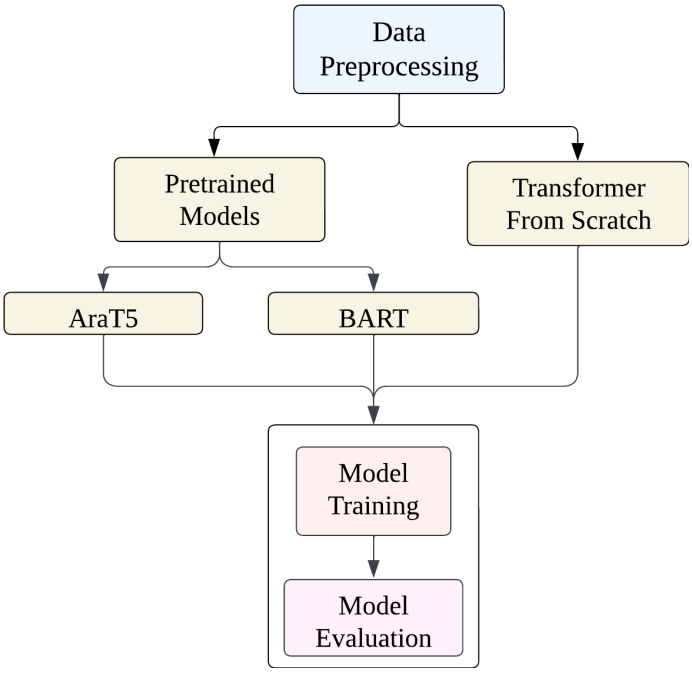


Fig. 1. Methodology

Islamic fatwas, each entry including a title, question, and corresponding answer. The dataset was made available on Kaggle [6] with the primary intention of aiding the research community.

In the data preprocessing stage, we undertook several key steps to ensure the dataset was clean and ready for analysis. Initially, we dropped all rows containing null values and removed any duplicate entries to maintain the integrity of the dataset. We then filtered the text to remove any characters that were not Arabic letters or English numerals, ensuring that the data was both relevant and consistent. Additionally, we decided to drop the title column and focus exclusively on the (question, answer) tuple, as these elements are most pertinent to the tasks at hand. This preprocessing ensured a streamlined and high-quality dataset for subsequent model training and evaluation.

B. Model Training

- **AraT5**
  The T5 model is designed to treat every NLP task as a text-to-text problem, meaning that both input and output are represented as text sequences. This unified approach allows the model to be applied to a wide range of tasks, such as translation, question answering, and classification, without needing task-specific architecture changes. The T5 model, short for ”Text-to-Text Transfer Transformer”, is based on the Transformer architecture. It uses an encoder-decoder structure, where the input sequence of tokens is first mapped to embeddings and passed through the encoder. The encoder consists of stacked ”blocks,” each containing a self-attention layer followed by a feed-forward network, with layer normalization and residual connections applied at each stage. Dropout is used at multiple points to prevent overfitting. The decoder mirrors the encoder but includes an additional standard attention mechanism to focus on specific parts of the input sequence [7].

- **AraBART**
  The AraBART (Bidirectional and Auto-Regressive Transformers) model is a sophisticated denoising autoencoder designed to handle a wide range of natural language processing tasks, including text generation, translation, and comprehension. Its structure is based on a sequence-to-sequence (seq2seq) framework, integrating a bidirectional encoder, similar to BERT, and an autoregressive left-to-right decoder, similar to GPT as shown in figure 2. This combination allows AraBART to effectively capture the full context of sentences through its encoder while generating fluent, coherent text with its decoder. The pretraining process of AraBART involves corrupting the input text using methods like random sentence shuffling and a unique in-filling scheme, where spans of text are replaced with a single mask token. The model is then trained to reconstruct the original text, enabling it to learn complex dependencies and robustly understand language structure [8].
  AraBART is a versatile and powerful NLP tool, thanks to its flexible noise addition during pretraining, which includes techniques like token masking, deletion, and sentence permutation. This flexibility allows AraBART to generalize well across various tasks and text types. Its combination of bidirectional and autoregressive ca-

pabilities makes it particularly effective for summarization, dialogue, and question answering. AraBART has achieved state-of-the-art performance in benchmarks, including significant improvements in ROUGE scores and outperforming traditional methods in machine translation, demonstrating its high performance in both generation and comprehension.

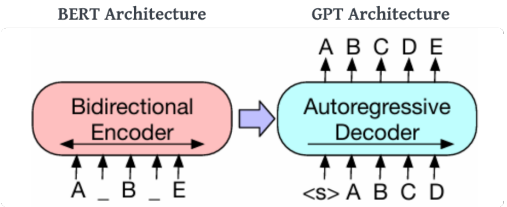


Fig. 2. AraBART Architecture

- **TransArab Transformer**
  The TransArab architecture is a breakthrough in sequence transduction, relying entirely on self-attention mechanisms without using recurrence or convolution. As depicted in Figure 3, it consists of an encoder-decoder structure, each composed of six identical layers. Each layer includes a multi-head self-attention mechanism followed by a position-wise feed-forward network, with residual connections and layer normalization to enhance learning. This architecture allows for highly parallelized computation, making the model both efficient in training and capable of achieving state-of-the-art performance in tasks like question answering.
  The TransArab model follows the architecture of the Transformer described in the paper *"Attention is All You Need"* [5]. A distinctive feature of this architecture is its multi-head attention mechanism, where eight parallel attention heads enable the model to focus on different parts of the input sequence simultaneously. Queries, keys, and values are projected into multiple subspaces, allowing diverse aspects of the sequence to be attended to in parallel. The model also uses sinusoidal positional encoding to capture the order of sequences and includes a masking mechanism in the decoder to ensure that each prediction only depends on previously generated outputs. These design choices make the Transformer a powerful and efficient tool for various sequence transduction tasks.

C. Evaluation Metrics

1) Precision: It measures the accuracy of the generated answer by evaluating how closely it matches the ground truth. It is calculated as the ratio of the number of common words between the generated answer and the ground truth to the total number of words in the generated answer

$$\text{Precision} = \frac{\text{Number of Common Words}}{\text{Number of Words in the Generated Answers}}$$

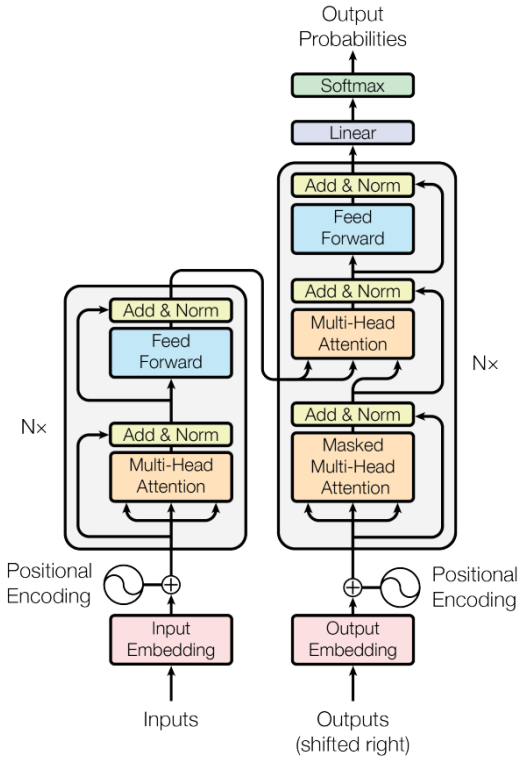


Fig. 3. Transformer Architecture

2) Recall: It determines how much of the relevant content (words) from the ground truth (the actual fatwa) is retrieved. Calculated as the ratio of the number of common words between the generated answer and the ground truth to the total number of words in the ground truth

$$\text{Recall} = \frac{\text{Number of Common Words}}{\text{Number of Words in the Ground Truth}}$$

3) F-measure: The F-measure, also known as the F1 score, is a metric that combines precision and recall into a single value to assess a model's overall performance.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4) BLEU Score: The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of text generated by a model, particularly in text generation tasks. It compares the n-grams of the generated text with those of a reference text. In our implementation, we use 4-grams to calculate the BLEU score. It is calculated as follows:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{4} w_n \log p_n\right)$$

where BP is the brevity penalty, w_n are the weights (usually uniform), and p_n is the precision of n-grams.

TABLE I
MODELS RESULT

Models	Precision	Recall	F1-Score	Bleu Score
Fine-Tuned AraT5	0.318	0.275	0.28	0.119
Unfine-Tuned AraT5	0.154	0.05	0.072	0.15
Fine-Tuned AraBART	0.370	0.285	0.302	0.126
Unfine-Tuned AraBART	0.188	0.097	0.012	0.002
TransArab Transformer	0.204	0.177	0.169	0.07

V. RESULTS AND DISCUSSION

The results presented in the Table I provide a comparative analysis of the performance of various models—both fine-tuned and non-fine-tuned (unfine-tuned)—as well as a custom-built Transformer model, TransArab. The evaluation metrics used include Precision, Recall, F1-Score, and BLEU Score, which collectively give a comprehensive understanding of the models' effectiveness in handling Arabic language tasks.

- **Fine-Tuned vs. Unfine-Tuned Models**
  The fine-tuned versions of AraT5 and AraBART significantly outperform their unfine-tuned counterparts across all metrics. This difference highlights the importance of fine-tuning for achieving better performance in specific tasks. For instance, Fine-Tuned AraBART achieves the highest Precision (0.370), Recall (0.285), F1-Score (0.302), and BLEU Score (0.126) among all models, indicating its superior ability to correctly generate accurate and relevant outputs.
  In contrast, the unfine-tuned versions of AraT5 and AraBART show considerably lower performance, with Unfine-Tuned AraBART having the lowest F1-Score (0.012) and BLEU Score (0.002). This suggests that models pre-trained on general corpora may not be as effective for specialized tasks without further fine-tuning on relevant datasets.

- **TransArab Transformer**
  The TransArab Transformer, trained from scratch following the "Attention is All You Need" methodology, performs moderately well, especially when compared to the unfine-tuned models. It achieves a Precision of 0.204, Recall of 0.177, F1-Score of 0.169, and BLEU Score of 0.07. While these results do not surpass those of the fine-tuned models, they are significantly better than those of the unfine-tuned versions. This indicates that although starting from scratch can be effective, it may require more extensive training or data to reach the performance levels of fine-tuned models.

- **BLEU Score Observations**
  Interestingly, the BLEU Score for the unfine-tuned AraT5 (0.15) is higher than that of the fine-tuned version (0.119). This anomaly suggests that while fine-tuning improves the overall precision and recall, it might lead to overfitting or may reduce the model's generalizability in generating outputs that match the reference texts. However, this increase in BLEU Score does not correspond to better performance in other metrics like F1-Score, which indicates that fine-tuning is still crucial for balanced model performance.

VI. CONCLUSION AND FUTURE WORK

The results clearly demonstrate that fine-tuning is essential for maximizing model performance on specific Arabic language tasks. Fine-Tuned AraBART emerges as the best-performing model, but the custom TransArab Transformer also shows promise, particularly when fine-tuning is not feasible. These findings underscore the need for continued refinement and training of models on domain-specific datasets to achieve optimal performance in Arabic Question Answering systems. For future work, we plan to test various LLM sizes to optimize performance and accuracy and expand the dataset scope to cover a wider range of Arabic question-answering fields.

REFERENCES

[1] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "AraT5: Text-to-text transformers for Arabic language generation". In: *arXiv preprint arXiv:2109.12068* (2021).

[2] Mohammad A Ateeq et al. "Arabic Narrative Question Answering (QA) Using Transformer Models". In: *IEEE Access* (2023).

[3] Wissam Antoun, Fady Baly, and Hazem Hajj. "AraGPT2: Pre-Trained Transformer for Arabic Language Generation". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Ed. by Nizar Habash et al. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 196–207. URL: https://aclanthology.org/2021.wanlp-1.21.

[4] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. "ARBERT & MARBERT: Deep bidirectional transformers for Arabic". In: *arXiv preprint arXiv:2101.01785* (2020).

[5] Ashish Vaswani. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

[6] Abdallah El-Saadany. *Fatawa: A dataset for Islamic Question and Answering Systems*. Kaggle. 2022. URL: https://www.kaggle.com/datasets/abdallahelsaadany/fatawa.

[7] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[8] Mike Lewis et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *arXiv preprint arXiv:1910.13461* (2019).