# HistoryQuest: Arabic Question Answering in Egyptian History with LLM Fine-Tuning and Transformer Models

Samaa Maged, Asmaa ElMaghraby, Ali Marzban,Esraa Negm, Mohamed Essawey, Amira Ahmed, Wael Gomaa
*Artificial Intelligence Program,*
*School of Information Technology and Computer Science (ITCS),*
*Nile University, Giza, Egypt*
{sa.maged, a.elmaghraby,a.marzban ,e.negm, m.abdelmaged,a.elsharaby, wabouzed}@nu.edu.eg

***Abstract*—Question answering (QA) in Egyptian history presents a unique and complex challenge for Arabic natural language processing (NLP). This study aims to explore and assess how large language models (LLMs) can enhance the accuracy and performance of Arabic question answering (QA), specifically in this domain. To conduct this investigation, we utilize two comprehensive datasets: the Arabic History-QA dataset and the Contextual Articles Dataset, which cover pivotal historical periods. We evaluate transformer-based models, including AraBERTv2, BERT-large-Arabic with Retrieval-Augmented Generation (RAG), fine-tuned LLaMa-2, and zero-shot LLaMa-3 with Retrieval-Augmented Generation (RAG). Through a rigorous and detailed evaluation process, we analyze how these models address various questions related to Egyptian history. This research contributes valuable insights into advancing the capabilities of Arabic NLP in specialized domains such as historical question answering. Our best results, summarized as the superiority of LLMs, beat those with transformers; additionally, the RAG significantly raised the performance level overall.**

***Index Terms*—Question Answering(QA),Arabic Question Answering(AQA) Large Language Models(LLMs), Natural Language Processing(NLP), Retrieval-Augmented Generation (RAG)**

## I. Introduction

Question Answering (QA) is a subfield of Natural Language Processing (NLP) that focuses on developing systems capable of automatically answering questions posed by humans in natural language [1]. Unlike traditional search engines that retrieve relevant documents, QA systems directly generate concise and accurate answers to user queries. A QA system typically consists of three main modules: question analysis, passage retrieval, and answer extraction [2]. The development of Arabic QA systems faces unique challenges due to the complexities of the Arabic language and the limited resources available for research. Arabic Question Answering (QAS) encounters linguistic complexities stemming from Arabic's rich morphology and intricate syntax [2].

Egyptian history holds profound significance in the domain of Question Answering (QA) due to its rich cultural heritage and historical importance. By collecting datasets like the Arabic History-QA dataset and the Contextual Articles Dataset, researchers can assess the effectiveness of QA systems in extracting accurate and contextually relevant information from Egyptian historical texts. This endeavor not only contributes to the advancement of Arabic NLP but also to historical knowledge related to Egypt and its pivotal role in shaping human civilization.

In this paper, we will propose techniques such as large language models (LLMs), retrieval-augmented generation (RAG), and transformer-based architectures. LLMs, pre-trained on vast amounts of text data, exhibit exceptional capabilities in understanding and generating natural language text. RAG combines retrieval-based and generation-based approaches, enabling QA systems to leverage external knowledge sources for more accurate and contextually relevant answers. Transformer-based models, including AraBERTv2 and BERT-large-Arabic with RAG, have demonstrated remarkable performance in various NLP tasks, including question answering.

By harnessing the potential of LLMs, RAG, and transformer models researchers can overcome the challenges inherent in Arabic Question Answering (AQA). These advanced technologies enable more effective passage retrieval, answer extraction, and overall performance improvement in Arabic Question Answering (QA). Our HistoryQuest: Arabic Question Answering in Egyptian History the system takes a question like في أي عام تولى محمد علي باشا الحكم في مصر؟ and receive the answer تولي محمد علي الحكم ١٨٠٥.

The rest of this paper is organized as follows: Section 2 presents Related Work, Section 3 shows the Methodology, Section 4 Results, and Section 5 concludes.

## II. Related Work

The question Answering (QA) field has attracted considerable interest in recent years. In our investigation of HistoryQuest: Arabic Question Answering in Egyptian History, we will overview the utilization of transformer-based models, large language models (LLMs), and Retrieval-Augmented Generation (RAG) techniques in Question Answering within the domain of Question answering. Additionally, we will provide an overview of the datasets and models employed in the reviewed papers, as shown in Figure 1.

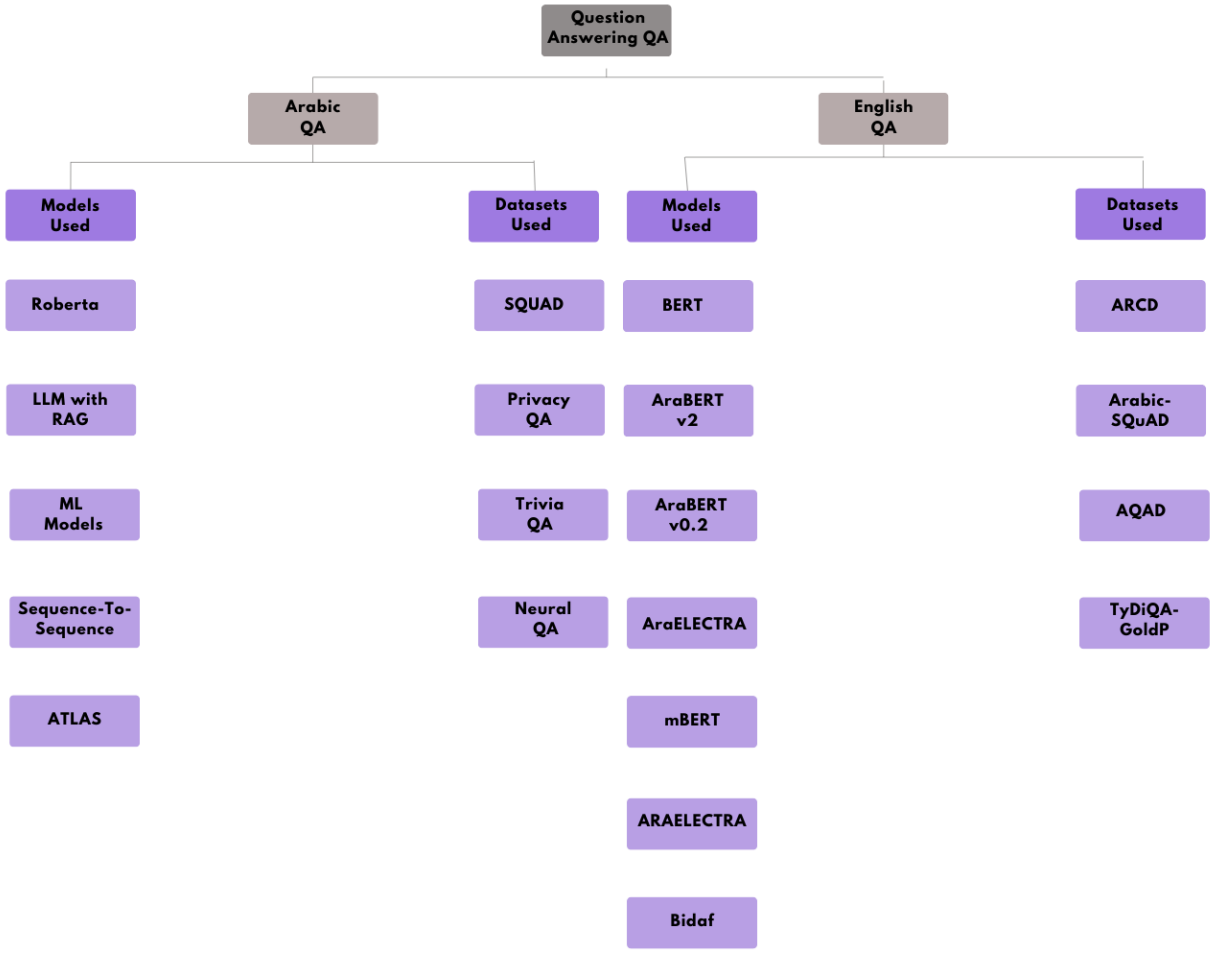


Fig. 1. Papers overview

### A. *English Question Answering*

Numerous approaches have been proposed in recent developments in English Question Answering. Yang and Ishfaq [3]introduced a multi-stage process that encoded context paragraphs at different levels of granularity, on the Stanford Question Answering Dataset (SQuAD) [4] that consists of 100,000+ questions posed by crowd workers on a set of Wikipedia articles. The best model achieved a 62.23% F1 score and a 48.72% EM score on the test set.Vold and Conrad [5] proposed RoBERTa Base question answering classification model in a production environment. Additionally, they compared the answer retrieval performance of a RoBERTa Base classifier against a traditional machine learning model in the legal domain on the publicly available PrivacyQA dataset [6] which consists of 1750 questions about the privacy policies of mobile applications and over 3500 expert annotations of relevant answers. Roberta achieved a 31% improvement in F1-score and a 41% improvement in mean reciprocal rank over the traditional SVM.

Izacard and Grave [7] proposed generative modeling and retrieval for open-domain question answering. This method involved two steps: first retrieving supporting passages using sparse or dense representations, and then using a sequence-to-sequence model to generate the answer based on the retrieved passages and the question. on the TriviaQA [8] dataset, which contains over 650000 question-answer-evidence triples and 95000 question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents.

Izacard et al. [9] presented Atlas, a carefully designed and pre-trained retrieval augmented language model capable of learning knowledge-intensive tasks with few training examples. Evaluations were conducted on MMLU [10], KILT [11], and NaturalQuestions [12].They also studied the impact of the content of the document index, demonstrating that it could be easily updated. Notably, Atlas achieved over 42% accuracy on natural questions using only 64 examples, outperforming a 540B parameter model by 3% despite having 50 times fewer parameters. Wang et al. [13]developed a new architecture for an LLM-based Retrieval-Augmented Generation (RAG) system by incorporating a specially designed rank head that precisely assessed the relevance of retrieved documents. Furthermore, they proposed an improved training method based on bi-granularity relevance fusion and noise-resistant training. By combining the improvements in both architecture and

training, they proposed that REAR was able to better utilize external knowledge by effectively perceiving the relevance of retrieved documents. Experiments were conducted on four open-domain QA tasks.

### B. Arabic Question Answering

Mozannar et al. [14] proposed an approach for open domain Arabic QA and introduced the Arabic Reading Comprehension Dataset (ARCD) Composed of 1,395 questions posed by crowd workers on Wikipedia articles, and a machine translation of the Stanford Question Answering Dataset (Arabic-SQuAD). The approach consisted of a document retriever using hierarchical TF-IDF and a document reader using BERT. They achieved an F1 score of 61.3% and a 90.0% sentence match on ARCD and a 27.6 % F1 score on an open domain version of ARCD. Alsubhi et al. [15] evaluated pre-trained transformer models for Arabic question answering using four reading comprehension datasets: ARCD [14], AQAD [16] is an Arabic reading comprehension large-sized high-quality dataset consisting of 17,000+ questions and answers, and TyDiQA-GoldP [17] which covers 11 typologically diverse languages with 204000 question-answer pairs. They fine-tuned and compared the performance of the AraBERTv2-base model, the AraBERTv0.2-large model, and the AraELECTRA model.

Atef et al. [16] fine-tuned the BERT-based un-normalized cased multilingual model(mBERT), which is trained on multiple languages, including Arabic. They also evaluated the model and used Arabic fastText embedding. They achieved 33 Exact Match(EM) and 37 F1-score and 32 EM and 32 F1-score using mBert and BiDAF model on AQAD dataset. Antoun et al. [18] developed ARAELECTRA, a pre-training method for Arabic language understanding. ARAELECTRA features a discriminator network with a BERT-like architecture and layers, for reading comprehension tasks. Their evaluation encompassed various Arabic NLP tasks, including reading comprehension, where the question answering task served as a measure of the model's reading comprehension and language understanding capabilities.

Compared to English language question answering systems, progress in Arabic question answering has been relatively slow due to limited resources and specific Arabic datasets. This has resulted in incomplete QA research, especially regarding large language models (LLMs) and LLMs with Retrieval-Augmented Generation (RAG) in Arabic. Our contribution to HistoryQuest: Arabic Question Answering in Egyptian History focuses on fine-tuning LLMs and using transformer models to improve Arabic QA. Comprehensive datasets like the Arabic History-QA dataset and the Contextual Articles Dataset are essential for advancing research in this domain.

## III. Methodology

Using the Arabic History QA dataset, which covers important periods of Egyptian history like the Pharaohs' era, the Roman occupation and Muhammad Ali's time, the Ptolemaic Dynasty, and the Greek era with Alexander the Great, our project aims to evaluate the efficiency of various text question answering models. Bu using five different ways to answer questions: the AraBERTV2 [15] transformer model, the BERT-large-Arabic transformer [19] with Retrieval-Augmented Generation (RAG) [20],LLaMa-2 [21]and zero-shot question answering using LLaMA-3 [22]. The compilation of the datasets, the architecture and operational framework of each model, and the procedures for training, fine-tuning, and evaluating them to compare their performance precisely are all covered in this methodology part.

### A. System Description

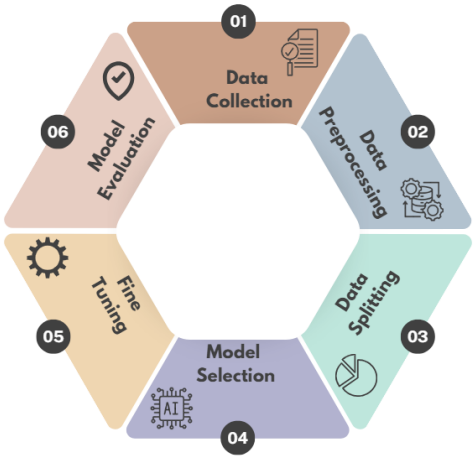


Fig. 2. The TQA Model Processing Pipeline.

The system used is described in two ways: first as shown in Figure 2, it involves data collection that was collected from Wikipedia and history articles also history books collected as questions and answers, prepossessing, Data splitting, fine-tuning the model and model evaluation, working as feeding an Arabic question into our model, which then produces an Arabic response to the question about Egyptian history.


Fig. 3. Data Distribution for Model Training.

As shown in Figure 3, Splitting Our data concluded at 95% for training, which represents 399 queries, and 5% for testing, which represents 21 queries of the data. The second system description is related to the use of the RAG part in Figure 4.First the data collection part web scrapped from Wikipedia articles into multiple articles in a PDF after that retrieving informative knowledge from the documents, choosing the most relevant one, and then utilizing the document and providing a prompt to produce an accurate response.

### *B. Datasets Description*

*1) HistoryQA Dataset:* The Arabic History QA dataset comprises 420 question-answer pairs, collected from historical texts, scholarly articles, and Wikipedia entries. It covers key Egyptian historical periods, including the Era of the Pharaohs, the Roman Occupation and Muhammad Ali's era, the Ptolemaic Dynasty, and the Greek period with Alexander the Great. This dataset is designed to test advanced Arabic natural language processing models.

*2) Contextual Articles Dataset:* The Contextual Articles dataset comprises a collection of articles in PDF format, which serve as a knowledge base for the Retrieval-Augmented Generation (RAG) component of our question-answering system. This dataset is specifically sourced from web-scraped articles covering eight distinct eras of Egyptian history. It provides additional historical details that enhance the answers produced by our model.

### *C. Models Description and Implementation*

*1) AraBERTv2 Transformer:* The AraBERTv2 model is a specialized adaptation of the well-known BERT architecture, specifically optimized to handle the complexities of the Arabic language. In this study, AraBERTv2 is utilized to analyze and interpret the History QA dataset with high precision. Preprocessing the data is a critical first step, involving the mapping of answers and context to fit specific model input requirements, ensuring that the data is in the optimal format for processing. AraBERTv2 is trained for 10 epochs with a batch size of 16 to fine-tune. Additionally, fine-tuning involves modifying several hyperparameters to strike a balance in the trade-off between model accuracy and training time.

*2) BERT-large-Arabic transformer model integrated with Retrieval-Augmented Generation (RAG):* We leverage a Transformer architecture equipped with a Retrieval-Augmented Generation (RAG) component to enhance the question answering capabilities of our system. This setup begins with indexing our historical articles dataset, which consists of PDF files that have been carefully split and cleaned. The splitting process involves dividing each article into manageable chunks using a specified chunk size and overlap, ensuring comprehensive coverage and continuity of information [20]. For the retrieval component, like in Figure 4 , we use an embedding model, specifically the BERT-large-Arabic model [19], to convert text chunks into dense vector representations. These embeddings serve as a basis for retrieving the most relevant chunks in response to a query. Following their retrieval, the correct chunks are processed to produce responses that make sense and are relevant to the context. Using the indexed content of our historical reference collection, this strategy guarantees accurate results. By using these methods, our system can generate responses that are firmly based on historical facts, accurately representing the range and complexity of the topics addressed.

*3) LLaMA-2 Fine-Tuning:* LLaMA-2 is part of the Large Language Model Meta AI (LLaMA) family of models, designed for robust performance across various natural language processing tasks. This model is optimized for speed and flexibility in handling diverse text data. We fine-tuned LLaMA-2 on The History QA dataset to adjust it to the complex requirements of historical text analysis [21]. The fine-tuning process included 10 epochs and strategic adjustments such as Low-Rank Adaptation (LoRA) parameters to 64 to enhance self-attention, and LoRA Alpha to 16 to control the learning rate, ensuring the model learns new data while retaining pre-trained knowledge. A LoRA Dropout of 0.1 is used to prevent over-fitting. Additionally, bits and bytes quantization is applied to improve efficiency, reducing the computational load without sacrificing accuracy. These measures equip LLaMA-2 to deliver precise, context-aware responses in Arabic, making it an integral tool for our research.

*4) LLaMA-3 Zero-Shot Learning:* LLaMA-3 represents a significant advancement in the LLaMA series, boasting state-of-the-art open-source capabilities with models of 8B and 70B parameters This next-generation model delivers exceptional performance across a broad spectrum of industry benchmarks and demonstrates enhanced reasoning, coding, and instruction-following abilities. It sets a new standard for large language models, combining the latest advancements in AI with a commitment to open access and community-driven innovation [22]. Using LLaMA-3 8B parameters for zero-shot learning, use its strong pre-trained architecture. We incorporate LoRA upgrades, with a LoRA Dropout of 0.1 to guarantee the model's robust generalization across various historical contexts. We set LoRA to 64 for enhanced self-attention and LoRA Alpha to 16 for optimized learning rate management. By improving its efficiency, bits and bytes quantization guarantees that LLaMA-3 performs well even when our computational resources are limited. By using this method, LLaMA-3 may demonstrate its adaptability to challenging question-answering circumstances.

*5) LLaMA-3 Zero-Shot Learning with Retrieval-Augmented Generation (RAG):* using the LLaMA-3 [22] model supplemented with Retrieval-Augmented Generation (RAG) [20] in our methodology. The Meta LLaMA-3 model, which has 8 billion parameters, is used because of its advanced zero-shot learning capabilities for handling complicated language tasks. This configuration consists of an extensive retrieval system based on our historical article dataset, which is first indexed and processed. The LLaMA-3 model uses preprocessing to split articles into manageable chunks, embedding them into dense vector representations using the BERT-large-Arabic model [19]. These embeddings are indexed using FAISS for ef-

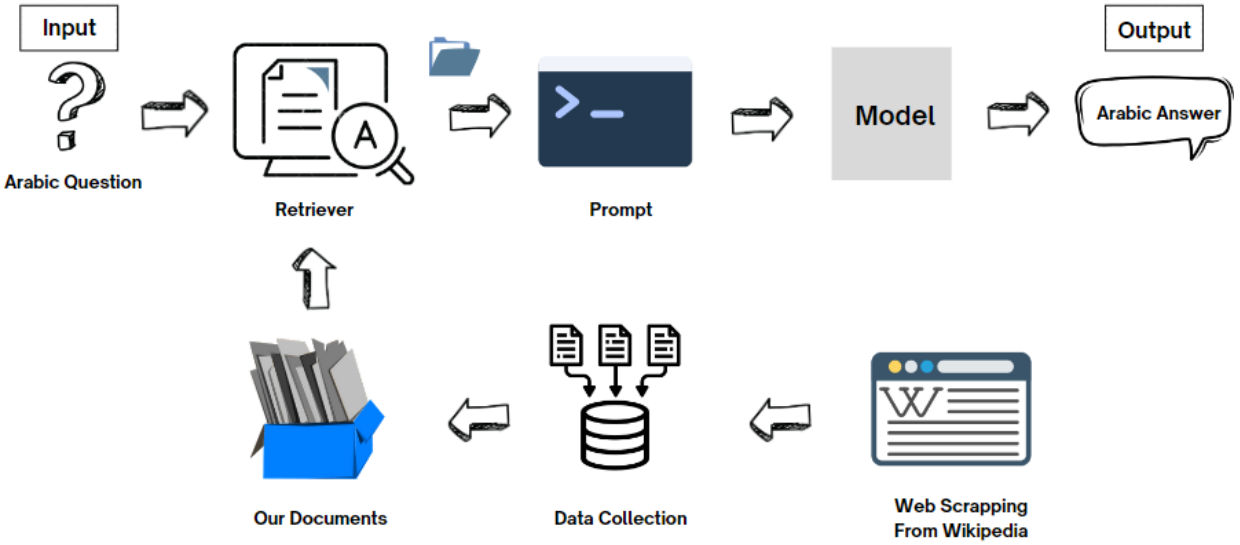


Fig. 4. The TQA using RAG Model Processing Pipeline.

TABLE I
MODEL'S RESULTS

| Model | Dataset | Loss | Correct Predictions | Incorrect Predictions |
|---|---|---|---|---|
| LLaMa-2 | History-QA | 0.2246 | 17 queries | 4 queries |
| LLaMa-3 | Zero-shot | - | 13 queries | 8 queries |
| LLaMa-3 with RAG | Contextual Articles Dataset | - | 18 queries | 3 queries |
| BERT-large-Arabic with RAG | Contextual Articles Dataset | 3.142 | 10 queries | 11 queries |
| AraBERT | History-QA | - | 7 queries | 14 queries |

ficient similarity search. The retrieval process fetches historical content for informed responses. The system is configured with parameters like LoRA attention dimensions and dropout rates to optimize performance and minimize overfitting. Techniques like bits and bytes quantization enhance model efficiency.

## IV. RESULTS

The performance of various Arabic Question Answering (QA) models is summarized in Table I. LLaMa-2 achieved a validation loss of 0.2246 after training on the History-QA dataset for 10 epochs, using a split of 399 examples for training and 21 for testing, as shown in Figure 3. LLaMa-3 was evaluated in a zero-shot setting, showcasing its performance without specific dataset fine-tuning. LLaMa-3 with RAG was tested on the Contextual Articles Dataset, although detailed results were not disclosed at this stage. AraBERT, trained for 10 epochs on the Contextual Articles Dataset, obtained a validation loss of 3.142. The evaluation involved binary classification, comparing predictions from ChatGPT-4 with our models' answers (LLaMa-3 zero-shot with RAG and transformer with RAG) and ground truth data for LLaMa-2 and AraBERT in the History-QA dataset. These findings highlight varying performance levels across models and datasets, emphasizing metrics such as correct and incorrect predictions to assess the efficacy of each model for Arabic QA tasks.

## V. DISCUSSION

AraBERT achieved 7 correct predictions and 14 incorrect predictions, highlighting challenges in handling historical context compared to LLaMa-2, which excelled with 17 correct predictions on the History-QA dataset. Conversely, BERT-large-Arabic with RAG, trained on the Contextual Articles Dataset, had 10 correct predictions out of 21, indicating that model size alone does not guarantee accuracy. LLaMa-3's zero-shot evaluation yielded 13 correct predictions and 8 incorrect predictions. Notably, LLaMa-3 with RAG demonstrated significant improvement in effectively answering queries from the contextual articles dataset, achieving 18 correct predictions out of 21. These findings underscore the critical role of tailored training and dataset selection in enhancing Arabic QA model performance for historical context understanding and information retrieval tasks. Transformer-based models like AraBERT and BERT-large-Arabic with RAG did not perform as well as Language Model Models (LLMs), especially when integrated with Retrieval-Augmented Generation (RAG), suggesting the potential superiority of LLMs in this domain. This underscores the need for further exploration of advanced fine-tuning methods and cross-lingual transfer learning to optimize model efficacy and generalizability in Arabic QA applications.

## VI. Conclusion and Future work

In conclusion, this study explored challenges and advancements in Arabic Question Answering (AQA) systems, focusing on Egyptian history. Using the Arabic History-QA dataset and the Contextual Articles Dataset, we evaluated transformer-based models like AraBERTv2, BERT-large-Arabic with RAG, fine-tuned LLaMa-2, and zero-shot LLaMa-3 with RAG, analyzing their performance in addressing diverse historical questions. Our findings demonstrate the efficacy of large language models (LLMs) in enhancing the accuracy and performance of Arabic QA systems. These advanced technologies are pivotal in overcoming the linguistic complexities inherent in Arabic, enabling systems to handle historical contexts with precision and nuance. In our future work, we aim to significantly enhance methods for Language Model Models (LLMs) and Retrieval-Augmented Generation (RAG) systems to gain deeper insights into model predictions, with a specific focus on understanding the underlying reasoning processes. We plan to explore cross-lingual transfer learning approaches to empower Arabic Question Answering (AQA) systems to effectively handle multilingual historical documents, thereby extending the applicability of these systems beyond single-language datasets. Additionally, we intend to refine and fine-tune LLaMa-3 using reinforcement learning with human feedback, leveraging iterative improvements based on real-world interactions to enhance model performance and adaptability. Another pivotal aspect of our upcoming research involves expanding the diversity and scale of datasets used for training and evaluation, with the goal of strengthening the robustness and generalization capabilities of our models across diverse contexts and domains.

## References


[1] M. A. C. Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 6, pp. 635–646, 2020.

[2] T. H. Alwaneen, A. M. Azmi, H. A. Aboalsamh, E. Cambria, and A. Hussain, "Arabic question answering system: a survey," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 207–253, 2022.

[3] C. Yang and H. Ishfaq, "Question answering on squad," *Department of Statistics*, pp. 1–7, 2018.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[5] A. Vold and J. G. Conrad, "Using transformers to improve answer retrieval for legal questions," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 245–249, 2021.

[6] A. Ravichander, A. W. Black, S. Wilson, T. Norton, and N. Sadeh, "Question answering for privacy policies: Combining computational and legal perspectives," *arXiv preprint arXiv:1911.00841*, 2019.

[7] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint arXiv:2007.01282*, 2020.

[8] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[9] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval augmented language models," *arXiv preprint arXiv:2208.03299*, 2022.

[10] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[11] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, *et al.*, "Kilt: a benchmark for knowledge intensive language tasks," *arXiv preprint arXiv:2009.02252*, 2020.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[13] Y. Wang, R. Ren, J. Li, W. X. Zhao, J. Liu, and J.-R. Wen, "Rear: A relevance-aware retrieval-augmented framework for open-domain question answering," *arXiv preprint arXiv:2402.17497*, 2024.

[14] H. Mozannar, K. E. Hajal, E. Maamary, and H. Hajj, "Neural arabic question answering," *arXiv preprint arXiv:1906.05394*, 2019.

[15] K. Alsubhi, A. Jamal, and A. Alhothali, "Pre-trained transformer-based approach for arabic question answering: A comparative study," *arXiv preprint arXiv:2111.05671*, 2021.

[16] A. Atef, B. Mattar, S. Sherif, E. Elrefai, and M. Torki, "Aqad: 17,000+ arabic questions for machine comprehension of text," in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–6, IEEE, 2020.

[17] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, "Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, 2020.

[18] W. Antoun, F. Baly, and H. Hajj, "Araelectra: Pre-training text discriminators for arabic language understanding," *arXiv preprint arXiv:2012.15516*, 2020.

[19] H. Chouikhi and F. Jarray, "Bert-based ensemble learning approach for sentiment analysis," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pp. 118–128, Springer, 2021.

[20] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[21] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[22] W. Huang, X. Ma, H. Qin, X. Zheng, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, and M. Magno, "How good are low-bit quantized llama3 models? an empirical study," *arXiv preprint arXiv:2404.14047*, 2024.