A Review of Question-Answering Systems Using Deep Learning in the Arabic Language

Ali Aloqla
Computer Information Systems Department, King Abdulaziz University, Jeddah, Saudi Arabia
aalaqla@kau.edu.sa (corresponding author)

Reda Khalifa
Computer Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia
rkhalifa@kau.edu.sa

Wajdi Alghamdi
Computer Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia
wmalghamdi@kau.edu.sa

Received: 21 August 2025 | Revised: 17 September 2025 and 20 September 2025 | Accepted: 24 September 2025

 DOI: https://doi.org/10.48084/etasr.14229

ABSTRACT

Question-Answering (QA) has become a pivotal topic in Natural Language Processing (NLP), facilitating machines' comprehension and response to human inquiries in natural language. Although QA systems for English and other high-resource languages have been extensively studied, Arabic QA remains under-investigated and faces several linguistic and technical challenges. This paper offers an extensive analysis of deep learning-based Arabic QA systems, emphasizing extractive, generative, and hybrid architectures. This study analyzes the fundamental issues in Arabic processing, outlines essential datasets, and provides a classification of QA methodologies. Furthermore, it identifies several research gaps, including the absence of domain-specific models, limited generative question answering, and insufficient use of retrieval-augmented architectures. To overcome these deficiencies, a Fatwa-based dataset, currently under development, can serve as a resource for future research on domain-specific Arabic QA. This study also delineates prospective trajectories, emphasizing the promise of Retrieval-Augmented Generation (RAG), few-shot learning, and dialect-aware models in propelling the discipline forward.

Keywords-Arabic NLP; QA; deep learning; RAG; natural language understanding; transformer

I. INTRODUCTION

Question Answering (QA) is a significant domain within Artificial Intelligence (AI) and Natural Language Processing (NLP) that aims to create computers proficient in comprehending and addressing human inquiries in natural language. With the increasing integration of digital assistants, customer care bots, and search engines into everyday life, the demand for efficient quality assurance systems has increased significantly.[1]. QA systems constructed using rule-based methods [2] or keyword-matching approaches [3] have shown deficiencies in semantic comprehension and contextual awareness. These systems were developed using Information Retrieval (IR) methods and Machine Learning (ML) techniques [4]. However, the true advancement occurred with the advent of Deep Learning (DL) models and transformer architectures, including Bidirectional Encoder Representations from Transformers (BERT) [5], Generative Pre-trained Transformer (GPT) [6], and Text-to-Text Transfer Transformer (T5) [7], which have markedly improved machines' capacity to understand and produce human language.

Foundational research in QA has broadly classified questions such as factual [8], definitional [8, 9], list-based [8], and causal queries [9]. QA systems frequently qualify as closed- or open-domain based on their domain scope [10, 11]. Closed-domain QA systems function within certain subject areas and frequently utilize structured knowledge bases or ontologies [10], whereas open-domain systems attempt to produce responses across a wide range of subjects utilizing unstructured corpora [11]. QA systems often integrate elements of IR, IE, and ML to derive succinct responses from extensive document repositories [12, 13]. The use of deep contextual embeddings has enabled these systems to more precisely align queries with pertinent response portions [5].

Despite considerable advances in high-resource languages such as English [14], the Arabic language continues to be markedly under-represented in QA research [15]. Arabic is one of the five most widely spoken languages worldwide; however, it poses distinct challenges such as intricate morphology, orthographic ambiguity, the absence of diacritics in most texts, and significant dialectal variance [16]. The language issues,

along with a lack of annotated corpora and restricted access to domain-specific information, provide significant challenges to the advancement of effective Arabic QA systems.

In the past decade, multiple initiatives have attempted to alleviate these challenges by creating Arabic QA datasets and pretrained language models. Datasets such as Arabic-SQuAD [17], QURAN-QA [18], and TyDiQA-Arabic [19] have facilitated further empirical investigations in the field. Arabic-specific pretrained models, such as AraBERT [20] and MARBERT [21], have established robust baselines across numerous downstream NLP tasks. Despite these advances, investigations into DL-based Arabic QA remain disjointed and lack thorough evaluations to inform future endeavors [17, 22]. This survey sought to synthesize existing research, pinpoint current issues, evaluate available tools and methods, and highlight interesting avenues for future research in Arabic QA utilizing DL.

A. Scope of the Survey

This review examines QA systems that employ DL methods for the processing and comprehension of inquiries in the Arabic language. The main objective is to evaluate the cutting-edge techniques and resources that facilitate Arabic QA. The focus was on systems engineered to respond to natural language inquiries utilizing neural architectures, encompassing Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models, such as BERT, T5, and GPT, along with Arabic-specific adaptations such as AraBERT and MARBERT. This study focuses on studies released between 2015 and 2025, a time frame that signifies the emergence and development of DL methods in QA. The survey includes several QA tasks, such as extractive, abstractive, multiple-choice, yes/no, and list forms. This study also examines several inquiry categories, including factual, definitional, causal (how/why), list-based, and hypothetical inquiries. Specifically, open- and closed-domain QA systems encompass a range of sectors, including academics, law, healthcare, and general knowledge. Multilingual and cross-lingual QA approaches are also examined for their evaluation or adaptation for Arabic. Considerable focus is directed toward publicly accessible Arabic QA datasets, such as Arabic-SQuAD, ARCD, TyDiQA-Arabic, and other benchmarks. The evaluation encompasses pretrained Arabic language models, including AraBERT, MARBERT, QARiB, and CAMeLBERT, with multilingual models such as mBERT and XLM-R. Common evaluation measures, such as Exact Match (EM), F1-score, BLEU, and ROUGE, are examined for Arabic QA evaluation. This study examines significant linguistic hurdles specific to Arabic, including intricate morphology, dialectal variation, discretization, tokenization problems, and orthographic ambiguity. Ultimately, it underscores the avenues for prospective studies, including short-term learning, cross-dialectal QA, and multimodal integration.

Although utilized as comparative baselines for Arabic, QA systems created specifically for English or other non-Arabic languages are not included in this survey to preserve its narrow focus. Systems that merely employ rule-based, template-based, or conventional information retrieval techniques without neural modeling are also excluded. Furthermore, non-automated QA systems that rely on human participation as a fundamental component are not taken into account, nor are chatbots and conversation agents until specifically assessed in QA settings. Specifically focusing on DL techniques used in Arabic QA, this review offers a comprehensive and methodical viewpoint on current developments, standards, difficulties, and potential research objectives in this rapidly expanding topic.

B. Contribution

There has been a dearth of in-depth research into the development of Arabic-specific QA systems based on DL, in contrast to the numerous developments in QA systems for high-resource languages. By compiling and analyzing the most recent studies in this field, this survey bridges this gap. This survey provides a focused investigation of Arabic QA systems, distinct from generic QA evaluations that primarily emphasize English or multilingual systems, and addresses the particular linguistic issues and resource constraints inherent in the Arabic language. The study also compares available Arabic QA datasets, highlighting their structure, question types, domain coverage, and suitability for training and assessing DL models. This evaluation of pretrained language models critically assesses Arabic and multilingual pretrained language models utilized in QA tasks, emphasizing their performance, limits, and adaptation methods. Subsequently, this investigation addresses open challenges, identifying critical issues such as dialectal variety, insufficient annotated resources, and inconsistencies in evaluation, and examines how they affect the efficacy of existing QA systems. Finally, the future roadmap delineates prospective avenues for the advancement of Arabic QA research, encompassing low-resource learning, cross-lingual transfer, and multimodal QA. This study provides a comprehensive and targeted evaluation that serves as a significant resource for academics, developers, and practitioners looking to develop or improve Arabic QA systems using DL-based technology.

II. FUNDAMENTALS

QA systems have been designed to deliver precise and succinct responses to inquiries presented in natural language. They exist at the convergence of various disciplines, including IR, NLP, and, to an increasing extent, DL.

A. Types of QA Systems

QA systems can be categorized into many types depending on their domain scope, answer generation technique, and reasoning complexity [1, 23, 24]. QA systems are primarily categorized as closed- and open-domain systems. Closed-domain QA functions within a defined field of expertise—such as medicine, law, or finance—and often depends on organized or curated information sources relevant to that domain [23]. In contrast, open-domain QA systems have been developed to address inquiries across diverse subjects, sourcing responses from extensive unstructured corpora such as Wikipedia or online articles [1, 23].

Methodologically, QA systems can be classified as extractive or abstractive [25]. Extractive QA identifies a segment of text directly from source documents that addresses

the question, rendering it especially appropriate for factoid-type inquiries and reading comprehension activities [25]. In contrast, abstractive QA seeks to provide replies in coherent natural language, frequently rephrasing or synthesizing information from several documents [26]. Certain QA assignments need intricate reasoning, such as multi-hop QA, where the system must synthesize and deduce information dispersed across several documents or evidence chains [26-28]. Retrieval-Augmented Generation (RAG) is a hybrid method that utilizes retrieval and generation, and its importance in answering Arabic questions is discussed in more detail in Section VI [29].

TABLE I. COMPARISON OF CLOSED-DOMAIN AND OPEN-DOMAIN QA

Aspect	Closed-domain QA	Open-domain QA
Scope	Specific subject area	General knowledge topics
Data source	Curated or structured corpus	Web-scale unstructured text
Accuracy	Typically higher	Typically lower
System complexity	Often simpler	More complex (retrieval + reasoning)
Use cases	Technical or academic support	Search engines, virtual assistants

B. QA Pipeline Architecture

QA systems often adhere to a multi-phase pipeline that amalgamates elements from IR, IE, and DL-based NLP [1, 26, 30]. This architecture's modular design allows systems to manage intricate queries, scale across domains, and include diverse models for varied activities [1, 31]. A conventional QA pipeline often has many interrelated components that convert a user query into an accurate and contextually relevant response [23, 26, 30]. The procedure starts with *question processing*, where the input query is subjected to several linguistic preprocessing stages, including tokenization, part-of-speech tagging, named entity identification, and question type categorization. In many systems, this phase further encompasses query reformulation to enhance retrieval accuracy [32]. Upon analyzing the question, the pipeline advances to *document retrieval*, whereby relevant passages are extracted from an extensive corpus utilizing either sparse techniques (e.g., BM25) or dense neural retrievers (e.g., Dense Passage Retrieval - DPR) [33]. This phase is regarded as the initial significant bottleneck in open-domain QA, since the quality of recovered information directly influences subsequent response creation or extraction [30, 33].

$$Score(q, p_i) = E_q^{\top} \cdot E_{pi} \quad (1)$$

where E_q represents the embedding vector of the query q, and E_{pi} signifies the embedding vector of the i-th passage p_i [33].

After retrieval, the system proceeds to the answer extraction or generation phase [29, 30]. Extractive models, exemplified by BERT, pinpoint a text segment from the retrieved documents that addresses the query [5], whereas abstractive models such as T5 or BART produce responses in coherent natural language by aggregating and paraphrasing information [7]. In certain implementations, an optional *answer ranking* phase is incorporated, wherein several potential responses are assessed using scoring heuristics or neural re-ranking models to identify the most contextually suitable answer [30]. This modular architecture facilitates adaptability across various QA paradigms and accommodates the incorporation of retrieval-augmented and generative methodologies [29, 34]. Figure 1 depicts a standard QA pipeline.

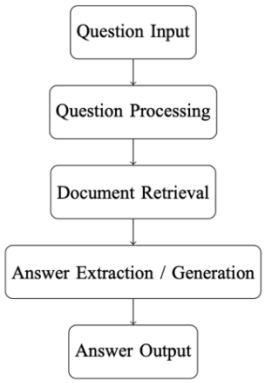

Fig. 1. Typical QA system pipeline.

In dense retrieval-based architectures, such as DPR and RAG, contrastive learning is used to efficiently train the retriever module. This is done by using a loss that maximizes similarity between the query and relevant passages while minimizing it for irrelevant passages. Contrastive loss is defined as:

$$\mathcal{L}_{retriever} = -log \frac{\exp\left(score(q,p^+)\right)}{\exp\left(score(q,p^+)\right)+\Sigma_{i=1}^{k} \exp\left(score(q,p_i^-)\right)} \quad (2)$$

For morphologically rich languages, such as Arabic, this formulation is vital as it enables the retriever to distinguish subtle semantic differences: q is the query embedding, p^+ the positive passage, p_i^- the negative passages, k the number of negatives, and $score(q,p)$ is the similarity function (e.g., dot product). This architecture has demonstrated efficacy in both closed- and open-domain QA systems [30], and it provides a versatile foundation for incorporating sophisticated modules such as RAG [29] and multi-hop reasoning [27].

III. EVALUATION METRICS

Evaluating QA systems requires metrics to assess the relevance, accuracy, and fluency of extracted or produced responses. One of the most used measures is Exact Match (EM), which assesses whether the predicted response is precisely aligned with the ground truth and is often used in datasets such as SQuAD [25]. The F1-score complements EM by calculating the harmonic mean of precision and recall at the token level, providing a nuanced assessment of responses that partially match the reference answers [25]. Precision is the percentage of predicted answer tokens that are correct, and recall is the percentage of ground truth answer tokens that were correctly predicted. The F1-score strikes a balance between these two metrics, being very useful for answers that may not be exactly right, but still show some correctness.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

The EM metric is more exacting. The projected response receives a score of 1 if it is exactly the same as the ground truth answer (after removing lowercasing, punctuation, and other irrelevant elements). If not, it receives a score of zero. This metric helps determine whether a response is entirely correct, and it is frequently used in conjunction with F1 to provide a more comprehensive assessment.

$$EM = 1_{\{A_{predicted} = A_{ground\ truth}\}} \quad (4)$$

In the field of generative QA, particularly for abstractive outputs, measures such as BLEU (Bilingual Evaluation Understudy) are commonly utilized. BLEU assesses n-gram correspondence between generated text and reference responses and was initially introduced for machine translation applications [35]. Likewise, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, especially ROUGE-L and ROUGE-N, focus on recall-based overlap and are widely utilized for summarization and generative QA evaluation [36]. In retrieval-based QA systems, Mean Reciprocal Rank (MRR) is commonly employed to assess the efficacy of the system in ranking the right response among a list of retrieved candidates [3].

In addition to automated metrics, human evaluation is essential in contexts requiring abstractive replies, where evaluators examine fluency, informativeness, grammaticality, and factual accuracy. Although inherently subjective, human assessment offers essential insights into response quality that automated scoring systems may neglect [37]. The selection of a suitable metric frequently relies on the type of QA operation (e.g., extractive vs. abstractive) and the system architecture. An effective evaluation method may integrate many automated measures with human assessment for a comprehensive evaluation [38, 39]. These essential elements and principles establish the basis for comprehending contemporary QA systems, particularly for their adaptation to morphologically intricate languages such as Arabic [40].

IV. ARABIC QA DATASETS AND BENCHMARKS

The advancement of Arabic QA systems has been continuously impeded by the lack of extensive, high-quality datasets. In the past few years, several annotated resources for both extractive and generative QA tasks have been released, differing in domain, question kinds, response formats, and linguistic scope (e.g., MSA vs. dialectal Arabic). The most important Arabic QA datasets are:

- Arabic-SQuAD [17]: An Arabic version of the Stanford QA Dataset (SQuAD), with more than 11,000 MSA questions, that adheres to the extractive QA structure and functions as a standard for Arabic reading comprehension assessments.
- ARCD (Arabic Reading Comprehension Dataset) [17]: Comprises more than 1,300 MSA QA pairs sourced from Wikipedia to evaluate reading comprehension in Arabic. It corresponds with SQuAD-style extractive configurations.
- TyDiQA-Arabic [19]: The Typologically Diverse QA benchmark encompasses Arabic along with ten other languages. It provides exemplary responses and texts, facilitating assessment in multi- and cross-lingual contexts.
- MADAR QA: Although not officially released, MADAR's dialectal corpora may be tailored for dialect-specific QA tasks, rendering it a valuable resource for extending beyond MSA [41].
- QALB (for paraphrase-driven QA): While mainly designed for grammatical correction, QALB's sentence-aligned structures may facilitate paraphrased QA matching in future QA assignments [42].

TABLE II. SUMMARY OF KEY ARABIC QA DATASETS

Dataset	Language Variety	QA Format	Domain
Arabic-SQuAD	MSA	Extractive	Wikipedia (general topics)
ARCD	MSA	Extractive	Wikipedia (general topics)
TyDiQA-Arabic	MSA	Extractive	Open-domain multilingual QA
QALB (adaptable)	MSA/Dialectal	Paraphrase (adaptable)	Educational error correction
MADAR (adaptable)	Dialectal	Multi-task (QA-capable)	City-level dialectal dialogue

V. LITERATURE REVIEW OF ARABIC QA SYSTEMS

Initial Arabic QA systems predominantly utilized rule-based methodologies and traditional information retrieval [43, 44]. These systems frequently employed pattern recognition, templates, or manually devised algorithms to discern question kinds and correlate them with pre-formulated responses [45]. For instance, in [46], a template-based system was developed to analyze input questions using meticulously built grammatical rules and align them with established answer templates. In [47], an information retrieval-driven QA system extracted pertinent paragraphs from texts by keyword overlap, subsequently identifying the most probable answer segment. These systems treated QA as a document retrieval issue rather than a genuine comprehension process [30].

These methods were straightforward to comprehend and computationally economical, although they exhibited inadequate generalization, restricted linguistic coverage, and a significant reliance on precise keyword matches. They were not resilient to morphological alterations or paraphrased inquiries and lacked the semantic reasoning abilities required for more intricate or implicit questions [1, 48, 49]. Certain systems, such as the one in [50], tried to integrate shallow parsing with ontological reasoning to improve the information retrieval process; however, it was constrained by limitations in rule coverage and scalability. These first endeavors established the foundation for Arabic QA but revealed that manually built and keyword-driven systems were inadequate to address the complete linguistic complexity and diversity of Arabic questions [40, 48].

A. Machine Learning and Feature-Based Models

As the constraints of rule-based QA systems became increasingly apparent, researchers began investigating statistical and ML methods for Arabic QA. These approaches often utilized supervised learning algorithms, including Support Vector Machines (SVMs), Decision Trees (DT), and Naïve Bayes (NB) classifiers, trained using explicitly produced

linguistic features [51]. Although these methods demonstrated considerable improvements in flexibility and generalization compared to rule-based systems, their reliance on manually crafted features constrained scalability and adaptability across many domains and languages [52]. In [53], an Arabic why-QA system (EWAQ) employed linguistic features and supervised classification methods to identify and extract appropriate response passages.

In this period, researchers showed an increasing interest in Arabic NLP preprocessing tools, including morphological analyzers and POS taggers, which became vital elements in conventional ML pipelines for QA and various NLP tasks. Technologies such as MADAMIRA [54] provide substantial assistance for morphological analysis and disambiguation in Modern Standard Arabic (MSA), facilitating the development of linguistic characteristics for subsequent tasks. However, despite their usefulness, these systems exhibited a deficiency in robustness and generalizability compared to contemporary end-to-end DL methodologies. In the lack of integrated learning across components, such pipelines frequently exhibited fragility and inadequate domain transferability [55].

B. Deep Learning-Based QA Models

The advent of DL marked a significant milestone in Arabic QA research. Models such as CNNs, RNNs, and attention mechanisms facilitate end-to-end learning of semantic patterns between questions and response passages by obviating the need for manual feature engineering [56]. In [57], cross-lingual language model pretraining was used to facilitate the response to questions in low-resource languages, such as Arabic. This method demonstrated that profound contextual comprehension, attained by multilingual embeddings, markedly enhanced passage rating and QA precision, even with minimal labeled data. Following the success of BERT [5], researchers began adapting transformer models to Arabic. AraBERT [20] was one of the initial monolingual Arabic transformers, specifically fine-tuned for extractive QA tasks such as Arabic-SQuAD [17]. Subsequent models such as MARBERT [21] and QARiB [21] focused on dialectal Arabic and domain adaptation, enhancing robustness for practical applications.

Regarding their achievements in extractive QA, most transformer-based systems were deficient in generative capabilities, as initial models were engineered for span extraction and categorization rather than text production [58]. This constraint prompted the creation of Sequence-to-Sequence (Seq2Seq) models, such as ArabicT5 [22], facilitating abstract QA and text production. However, these models are rarely integrated with retrieval processes, constraining their factual precision and making them susceptible to hallucination, particularly in knowledge-intensive activities [29]. Despite substantial progress in Arabic QA using DL, difficulties persist in tailoring models to the language's complex morphology, dialectal variation, and resource-scarce subdomains such as religious or legal literature [40]. Furthermore, many current systems are extractive and fail to meet the demand for explainable, retrieval-augmented, generative QA, highlighting the necessity for RAG-based approaches [26, 29]. Since most current systems remain primarily extractive, they have limited integration of retrieval and generative functionalities [59].

C. Multilingual and Cross-lingual QA

Due to the scarcity of extensive Arabic QA datasets and pretrained models in previous years, numerous studies resorted to multilingual models and cross-lingual transfer learning to develop QA systems for Arabic [60, 61]. These methods used multilingual transformers such as mBERT [5], XLM-R [61], and mT5 [62] to transfer knowledge from high-resource languages (mainly English) to low-resource ones such as Arabic. TyDiQA [63] was one of the initial multilingual QA datasets to offer a high-quality Arabic subset, allowing researchers to optimize multilingual models specifically for Arabic QA tasks. Although multilingual models such as mBERT and XLM-R provide wide coverage and zero-shot QA capabilities, they often fall short compared to monolingual models such as AraBERT due to inadequate tokenization and limited Arabic-specific pretraining [20, 64]. Researchers employed cross-lingual QA pipelines that translated Arabic inquiries into English, applied QA using English models, and then translated the responses back [65]. This strategy, although conceptually attractive, results in translation inaccuracies, meaning shifts, and loss of subtlety, which is particularly concerning for Arabic due to its complex morphology and cultural background [64, 66].

Although multilingual models facilitate extensive language coverage and provide zero-shot QA capabilities, they often do not adequately capture the intricate syntax, morphology, and dialectal variety characteristic of the Arabic language [20, 65]. These constraints arise from inadequate tokenization and insufficient pretraining tailored to Arabic in models such as mBERT and XLM-R. Moreover, as previously mentioned, most Arabic QA systems are still extractive, with minimal integration of retrieval and generation [59], which limits their efficacy in knowledge-intensive tasks where contextual grounding and response fluency are crucial [59, 65]. RAG presents a compelling alternative by integrating non-parametric memory (document retrieval) with generative modeling, therefore mitigating several deficiencies [29].

D. Domain-Specific Arabic QA

Although most Arabic QA research has focused on open-domain tasks using broad corpora, including Wikipedia, domain-specific QA is considerably underexplored. Domains such as religion, law, healthcare, and education require customized models and meticulously selected datasets due to the intricate and sensitive nature of the information involved [63, 67-70]. Preliminary research examined technical fields using structured ontologies [71]. In [72], a QA system was proposed for religious inquiries using hadith ontologies. In recent years, in [18], a benchmark dataset and a shared task focused on Qur'anic QA, assessing both passage retrieval and response production in Classical Arabic. In [73], Large Language Models (LLMs), such as GPT and BART, were used for Qur'anic QA, achieving robust generating outcomes. However, this study was confined to the Qur'anic corpus and lacked external retrieval mechanisms, prompting inquiries regarding factual foundation and transparency [18]. Despite these advances, a significant gap remains in QA research, addressing the comprehensive range of religious fatwas, which need subject expertise, language subtleties, and source citation

(e.g., Quran, Hadith, scholarly consensus). This is especially crucial in Islamic contexts, where precise sourcing and interpretation have substantial implications [74].

In [75], the incorporation of structured knowledge into LLMs for Quranic QA was investigated. The proposed configuration integrated knowledge triples and verse contexts into GPT-4 to address fact-based inquiries sourced from the Quran. The knowledge-infused variation demonstrated enhanced factual correctness, completeness, and interpretability compared to GPT-4 alone. Manual assessments verified that organized inputs significantly reduced hallucinations and provided more grounded responses. Although this method did not employ explicit retrieval techniques such as RAG, it was observed that knowledge triples operate comparably to a streamlined retrieval layer. This hybrid configuration reconciles generative fluency with factual accuracy, rendering it especially advantageous for delicate religious contexts.

Recent initiatives aimed to tackle domain-specific QA in Classical Arabic, especially concerning Islamic texts such as the Holy Qur'an. In [76], the ARAELECTRA model was meticulously refined using several Arabic datasets, including Ar-TyDi QA, Arabic-SQuAD, ARCD, and QRCD, to develop a transformer model for Qur'anic language. The model attained a partial Reciprocal Rank (pRR) of 66.9% on the training set and 54.59% on the test set. This study also examined various loss functions, including Focal Loss and Dice Loss, to address dataset imbalance. This tiered fine-tuning strategy demonstrated potential for adapting contemporary structures to historic religious texts, addressing constraints in dataset magnitude and language difference between MSA and Classical Arabic. This approach may be applied to more sensitive topics necessitating source attribution and semantic accuracy.

Recent studies have investigated domain-specific QA applications in Arabic, such as the HistoryQuest system designed for Egyptian historical materials [77]. This study presented the Arabic History-QA and Contextual Articles datasets and assessed several models, including AraBERTv2, BERT-large-Arabic with RAG, fine-tuned LLaMA-2, and zero-shot LLaMA-3. LLaMA-3, when combined with RAG, attained superior accuracy, surpassing conventional transformers by adeptly merging dense retrieval (utilizing FAISS indexing and BERT embeddings) with generative modelling. These findings underscore the efficacy of hybrid methods in resource-constrained and linguistically intricate environments such as Arabic QA.

To rectify this deficiency, this study proposes the development of a domain-specific Arabic QA system focused on Fatwas in the future. An innovative Fatwa QA dataset can curate Fatwa collections from verified religious portals in Saudi Arabia and the MENA countries. This dataset can provide QA using RAG, providing precise and elucidated responses based on reliable religious sources.

E. Morphological Complexity

Arabic is a morphologically complex language capable of producing numerous word forms through inflection, derivation, and clitic attachment [16]. A single lemma can manifest on several surfaces according to gender, number, tense, and case [16]. This exacerbates data sparsity and complicates the models' ability to acquire consistent representations unless adequately normalized [11]. These problems are especially evident in applications such as QA and document retrieval, where precise token matching is essential. To alleviate this issue, preprocessing approaches such as morphological segmentation, lemmatization, and clitic separation have emerged as vital elements of contemporary Arabic NLP pipelines [78]. Furthermore, contemporary language models such as AraBERT and MARBERT utilize pretraining on segmented or morphologically-informed corpora to more effectively capture the linguistic subtleties of Arabic [20, 21]. Despite these advances, effective management of morphological variation continues to be a significant challenge, particularly in dialectal or domain-specific settings [16].

F. Orthographic Variability and Diacritics

Most Arabic writings are composed without diacritics, which denote short vowels and additional phonetic indicators [16]. The lack of diacritics creates considerable uncertainty in reading and interpretation [16]. For instance, the word كتب (ktb) may mean "wrote," "books," or "was written," contingent upon the context [79]. Moreover, discrepancies in the representation of long vowels (e.g., ا), hamza (e.g., ء), and ta marbūta (e.g., ة) exacerbate the challenges of tokenization and retrieval [16, 17].

G. Dialectal Diversity

MSA is employed in official settings; however, a significant amount of real-world data—particularly on social media, conversations, and user-generated content—comprises regional dialects (e.g., Egyptian, Levantine, Gulf) [80]. These dialects exhibit considerable divergence in vocabulary, grammar, and spelling, and are not adequately covered in current NLP resources [41]. Informal Arabic writing frequently exhibits non-standardized spelling and may be rendered in Romanized script (generally known as Arabizi), thus complicating tokenization and morphological analysis [81, 82]. Dialectal Arabic has data scarcity, as most existing corpora and pretrained models are skewed towards MSA [83]. This discrepancy often leads to diminished model efficacy when utilized in real-world or user-generated texts [84]. Recent initiatives, such as the MADAR corpus [41] and models such as MARBERT [21], have begun to address these deficiencies by providing dialectal benchmarks and embeddings developed from informal Arabic. However, several dialects are still underrepresented, and comprehensive generative or retrieval-augmented models designed for these types are currently absent [84]. Filling this gap is crucial for the implementation of effective Arabic QA systems that can generalize across both formal and informal text [63].

H. Lack of Annotated Resources

Compared to English, Arabic experiences a deficiency of high-quality, extensive annotated datasets for QA and associated natural language comprehension tasks. Although initiatives such as Arabic-SQuAD and ARCD have provided significant resources to the community, they are constrained in terms of both scale and domain variety [17]. These datasets frequently emphasize Wikipedia-style material while

neglecting specialist domains such as religion, law, and healthcare, which are essential for practical QA applications [21]. Furthermore, the absence of comprehensive datasets for dialectal Arabic further limits the generalizability of trained models [41]. Initiatives such as TyDiQA seek to address this multilingual disparity; nevertheless, Arabic is still inadequately represented in both volume and task-specific profundity [63]. Mitigating these constraints is crucial for the advancement of precise and domain-adaptable Arabic QA systems [63].

I. Tokenization and Segmentation Issues

Arabic tokenization is complex due to the existence of connected clitics (such as conjunctions, prepositions, and pronouns) and the absence of whitespace segmentation for certain affixes. Inadequate tokenization can substantially impair retrieval and model precision [20]. To address these issues, morphological analyzers and tokenizers such as MADAMIRA [54] and Farasa [85] have been developed to enhance segmentation and disambiguation accuracy. These tools assist in discerning the accurate morphological forms of words and isolating clitics from base stems, which is crucial for subsequent QA performance [54, 85]. However, standardizing tokenization across languages and fields continues to pose a significant research problem.

J. Cross-dialectal and Multilingual Transfer Limitations

Although the robust performance of multilingual pretrained models such as mBERT and XLM-R across several languages, their efficacy in Arabic, especially in dialectal variations, frequently falls short. This is mainly attributable to the models' training on corpora mostly including MSA and insufficient dialectal representation, which constrains generalization to informal or regionally specific use [20, 21]. Furthermore, cross-lingual transfer from English to Arabic is hindered by significant syntactic, morphological, and lexical variation between the two languages, impacting semantic alignment in multilingual QA tasks [61]. These constraints highlight the necessity of developing Arabic-specific or dialect-sensitive models that can more effectively accommodate the linguistic variety inherent in the Arabic language [20].

In addition, dialects including Egyptian, Levantine, and Gulf Arabic are rarely included in the pretraining corpora, resulting in a discrepancy between model anticipations and actual user inputs in applications such as virtual assistants and social media QA. Multilingual models, although fine-tuned with Arabic data, encounter difficulties with details such as codeswitching, clitic attachment, and diacritic ambiguity, which considerably impact semantic comprehension [86]. Language and resource-related issues underscore the need for specific preparation tools, dedicated datasets, and customized architectures for Arabic QA. They also emphasize the possibility of modular hybrid frameworks like RAG to utilize retrieval for disambiguation and domain adaptation [29]. By separating retrieval from production, RAG allows independent fine-tuning of Arabic religious or dialectal corpora, thereby providing more flexibility and increased factual accuracy.

TABLE III. COMPARISON OF RECENT ARABIC QA STUDIES

Study	Approach	Language	Domain	Limitations
AraQA-BERT [87]	Extractive (AraBERT)	MSA	Open-domain	Lacks generative capabilities; not domain-specific
ArabicaQA [63]	Dataset + Dense Retrieval (AraDPR)	MSA	Open-domain	Lacks domain-specific questions; no generative QA
Qur'an QA [18]	Passage Retrieval + MRC	Classical Arabic	Religious (Qur'an)	Focused on extractive QA; limited to Qur'anic text
LLMs for Qur'anic QA [73]	Generative (GPT-4, BART)	Classical Arabic	Religious (Qur'an)	High performance but limited to Qur'anic domain; lacks retrieval grounding
MQA-KEAL [88]	Multi-hop QA with Knowledge Editing	MSA	Open-domain	Focused on knowledge editing; not tailored for religious texts

K. Challenges in Arabic QA

Implementing efficient QA systems for Arabic entails distinct challenges, mostly resulting from the language's intricate morphology, orthographic intricacies, and dialectal diversity. These variables render the direct use of approaches effective in English or other high-resource languages inadequate for Arabic QA tasks [16, 89].

VI. QA MODELS

A. Extractive vs. Abstractive QA

QA systems may be classified into extractive, abstractive, and hybrid methods. Extractive QA systems pinpoint and retrieve a segment of text from a specified context that directly addresses the inquiry [26]. These models, frequently founded on transformer encoders such as BERT [5] or AraBERT [20], are extensively used in response to their interpretability and robust performance on benchmarks such as SQuAD [25] and Arabic-SQuAD [17].

In contrast, abstractive QA systems produce responses in natural language that may not precisely correspond to any segment inside the context. These systems utilize Seq2Seq models such as T5 [7], mBART [90], or ArabicT5 [22]. Although they provide natural, human-like responses and multi-document synthesis, they are more challenging to assess and may generate unsubstantiated material. To integrate the advantages of both paradigms, RAG has developed as a hybrid architecture [29]. RAG acquires pertinent documents and transmits them to a generative model that generates contextually informed responses [29]. This design enhances factual accuracy and facilitates knowledge-intensive tasks [29].

Figure 2 delineates the structural distinctions among the three QA paradigms. Extractive systems execute direct span selection [5], abstractive systems formulate answers from the input text [26], and RAG models retrieve and conditionally create responses utilizing external documents [29]. The hybrid

technique shows significant potential for Arabic QA in specialized fields, such as religion or medicine, where accurate and generative solutions are essential [20, 29, 40]. For the question "Who is the founder of Microsoft?", an extractive answer would be "Bill Gates" (extracted directly from the source text), and an abstractive answer would be like "Microsoft was co-founded by Bill Gates and Paul Allen."

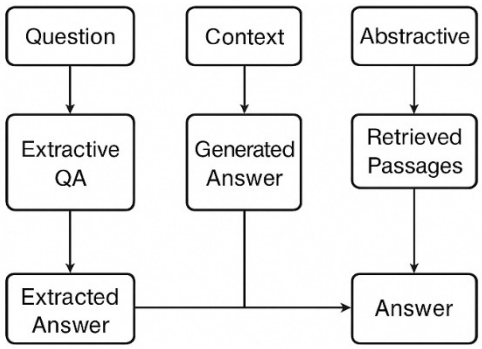

Fig. 2. Comparison of Extractive, Abstractive, and RAG-based QA pipelines, adapted from [26, 29].

TABLE IV. COMPARISON OF QA APPROACHES

Feature	Extractive QA	Abstractive QA	RAG (Hybrid)
Answer source	Span from input text	Generated from context	Generated from retrieved documents
Model type	Encoder only	Encoder-Decoder	Retriever + Generator
Fluency	Low to moderate	High	High
Factuality	High (if input is reliable)	May hallucinate	High (if retrieval is good)
Interpretability	High	Medium	Medium to High
Data requirement	Moderate	High	High (documents + questions)
Arabic models	AraBERT, MARBERT	ArabicT5, mBART	RAG with AraBERT + ArabicT5

B. Retrieval Augmented Generation

RAG is a hybrid neural architecture that incorporates information retrieval into the language model generation process [29]. It was intended to augment the functionalities of extensive generative models in knowledge-intensive activities, including open-domain QA, fact verification, and summarization [29]. In contrast to conventional generative models such as GPT or BART, which operate purely on the information encoded in their pretrained parameters (i.e., parametric memory), RAG employs a hybrid method that addresses this intrinsic restriction. As existing models are unable to dynamically assimilate new or domain-specific knowledge without undergoing retraining or fine-tuning, RAG enhances their functionality by including a non-parametric memory—specifically, a dense document retrieval mechanism—alongside a generative model. This architecture facilitates real-time access to external information sources, thus improving flexibility and factual accuracy in knowledge-intensive tasks [29].

1) RAG Architecture

The RAG architecture consists of two primary, tightly integrated modules that work in tandem to enhance the accuracy and contextual richness of generated answers [29]. First, the retriever component—typically implemented using a dense retrieval model such as DPR—encodes the input question and retrieves the top-k most relevant documents from a large corpus, which may include Wikipedia, structured knowledge bases, or domain-specific text collections [33]. Next, the generator module, usually based on powerful Seq2Seq models like BART or T5, processes each retrieved document concatenated with the original question to generate candidate answers [7, 58]. These generated responses are then scored and aggregated to produce the final answer. By following this retrieval-then-generation paradigm, RAG systems can dynamically incorporate external knowledge at inference time [29]. This makes them more interpretable and adaptable than purely generative models, especially in knowledge-intensive QA tasks [91].

$$P(y \mid x) = \sum_{d \in \mathcal{D}(x)} P(y \mid x, d) \cdot P(d \mid x) \quad (5)$$

where x signifies the input question, and d $\in \mathcal{D}(x)$ indicates the top-k documents extracted from the corpus in response to the query. The notation $P(d \mid x)$ represents the retriever's assessed probability of document d being pertinent to the input x, whereas $P(y|x, d)$ denotes the probability of producing the response y based on both the input question and the retrieved document [29].

2) RAG in QA

In the field of QA, RAG functions as a proficient intermediary between extractive and abstractive methods [29]. In contrast to extractive approaches, such as BERT-based span selectors, which only locate and reproduce segments of the source information, RAG may provide coherent, human-like answers that transcend simple span replication. In contrast to conventional generative models that depend exclusively on their internal parametric memory, RAG bases its responses on retrieved texts [58]. This grounding enhances factual consistency and increases the system's robustness in low-resource or domain-specific contexts. Consequently, RAG is especially adept at open-domain QA tasks, where the model must generate coherent and contextually relevant responses from a vast and varied corpus [29].

VII. DL APPROACHES FOR ARABIC QA

DL has transformed the domain of QA by facilitating end-to-end learning and markedly enhancing the efficacy of both extractive and generative QA models. This section delineates the principal neural architectures utilized in Arabic QA systems, encompassing both generic and Arabic-specific pretrained models.

A. Neural Architectures in QA

Initial DL methods in QA included CNNs [92] and RNNs [93] for the representation of sentences and the alignment of question-context pairings. LSTMs were developed to capture long-range relationships and increase sequential modeling [94], hence improving span prediction and passage ranking in QA

processes. Despite their early success, these models were limited in their ability to capture global context and encountered difficulties with intricate syntactic relationships, resulting in the adoption of self-attention-based models.

B. Transformer-Based Models

The advent of the transformer architecture [95] represented a substantial progression in NLP, particularly in QA. Transformers facilitated parallel processing and improved contextual comprehension by substituting recurrence with self-attention processes. BERT [5] utilized bidirectional attention to generate comprehensive contextual embeddings, establishing itself as a fundamental component for several extractive QA systems. Several transformer-based models have been particularly adapted for the Arabic language, building on this basis. AraBERT [20] modified the BERT architecture by pretraining on an extensive Arabic corpus comprising Wikipedia, news articles, and the OSIAN corpus, demonstrating robust performance on several downstream tasks, including Arabic QA. Subsequently, MARBERT [21] augmented this research by training on dialectal and informal Arabic material sourced from Twitter, therefore improving QA performance on social media and practical contexts. Similarly, QARiB [21] was developed utilizing varied datasets to improve representation in both MSA and dialects, thus providing more applicability in QA tasks. Multilingual models such as mBERT and XLM-R [57] have been utilized in Arabic QA. Their training on more than 100 languages, including Arabic, establishes them as robust baselines for multilingual and cross-lingual applications, while not being language-specific. In summary, these models illustrate the increasing sophistication of Arabic-specific transformer topologies and their essential contribution to enhancing QA capabilities for the Arabic language.

C. Generative Models for QA

Unlike extractive models that depend on distinguishing text segments from a particular context, generative models can generate free-form responses in natural language [6, 7]. These models utilize transformer-based Seq2Seq topologies to produce coherent and contextually pertinent replies. Prominent instances comprise T5, which conceptualizes all NLP tasks as text-to-text challenges, facilitating unified generative modeling [7], and GPT-3, which exhibits few-shot and zero-shot proficiency across diverse QA contexts through extensive pretraining [6]. Although specialized Arabic generative models are currently being developed, advances have been achieved through multilingual pretrained models. Specifically, mT5 [62] and mBART [96] have demonstrated robust performance when fine-tuned for Arabic tasks, utilizing extensive multilingual datasets and facilitating text-to-text transformations, rendering them appropriate for generative QA in Arabic. In the absence of an officially released Arabic-specific T5 model, these multilingual architectures continue to serve as robust baselines and transfer learning frameworks in Arabic QA research [90]. These improvements indicate an increasing interest in creating fluent and semantically robust QA systems for Arabic, transitioning from extractive techniques to harnessing the capabilities of generative architecture [97].

D. Retrieval-Augmented Generation

RAG [29] is a hybrid model that integrates dense document retrieval with neural generation. It mitigates a fundamental constraint of exclusively generating models—the lack of access to external knowledge—by retrieving pertinent sections and conditioning a generator on both the query and the acquired material [29]. In Arabic QA, RAG presents a potential avenue, particularly for specialized applications such as religious domain retrieval [18, 22]. The distinction between retriever and generator facilitates adaptable configurations: a retriever may be specifically refined using Arabic religious corpora, whilst a multilingual generator (e.g., mBART or ArabicT5) generates contextually relevant and fluent replies [21, 29].

Most recent advances in RAG systems have sought to connect language models with external structured information sources, such as Knowledge Graphs (KGs). However, KG-based RAG pipelines often underperform in intricate reasoning tasks. Mindful-RAG [98] is a restructured KG-RAG framework that explicitly addresses eight persistent points of failure, including reasoning misalignments and structural deficiencies. This method integrates LLM-driven intent identification, contextual constraint management, and validation procedures to ensure that responses correspond to user inquiries. Experimental findings on the WebQSP and MetaQA datasets demonstrate significant improvements, particularly in minimizing hallucinations and enhancing multi-hop reasoning accuracy. The findings indicate that forthcoming Arabic QA systems may greatly benefit from the incorporation of intent-aware modules and constraint-driven retrieval techniques similar to Mindful-RAG, especially in knowledge-intensive areas such as Fatwa and Islamic law.

A notable recent contribution is the RFPG framework, which improves the RAG architecture by including a fact-checking layer in the retriever component to boost factual correctness in low-resource contexts, particularly in Arabic QA tasks [99]. RFPG focuses on the hadith domain, utilizing a meticulously selected corpus of 34,542 genuine hadiths. The pipeline has four primary stages: Retrieval of Hadiths, Verification of Facts, Engineering of Prompts, and Generation of Responses. This design greatly reduces hallucinations and increases reliability. This study demonstrated the superiority of RFPG compared to traditional RAG- and GPT-based models (GPT-4, GPT-4o, GPT-4o-mini) on 123 Arabic inquiries. The framework achieved a 100% accuracy and 98% precision in hadith references, surpassing GPT-4o-mini, which fabricated 42 out of 123 responses. The model's efficacy was attributed to its specialized tuning and retrieval basis, rendering it especially appropriate for applications related to religious texts or critical factual QA. RFPG exemplifies how domain-adapted, fact-aware RAG designs may mitigate hallucination and generalization challenges in Arabic QA systems, particularly when engaging with intricate or religious literature.

E. Summary and Observations

DL has facilitated considerable advances in Arabic QA; yet, most models continue to rely on architectures initially designed for English. Although Arabic-specific models such as AraBERT and MARBERT investigate certain linguistic subtleties, difficulties persist in managing dialects, domain-

specific terminology, and low-resource environments. Hybrid methods such as RAG provide new avenues for resilient, interpretable, and precise QA in Arabic.

Although designs such as BERT and T5 exhibit robust performance, their implementation in Arabic encounters distinct hurdles stemming from the language's intricate morphology and dialectal diversity [20]. In response to these issues, models tailored for Arabic, such as AraBERT [20] and MARBERT [21], have been created. AraBERT is pretrained on extensive MSA corpora, whereas MARBERT emphasizes dialectal Arabic and social media material, hence augmenting model resilience across linguistic variants. Despite these advances, several models remain inadequate when used in domain-specific or low-resource settings [100, 101]. In response, hybrid designs such as RAG [29] present a possible option. By separating the retrieval and generation processes, RAG facilitates more interpretable, precise, and contextually aware QA, which makes it especially appropriate for Arabic domains that need both linguistic subtleties and access to external knowledge [29, 100].

VIII. DOMAIN-SPECIFIC ARABIC QA

To enhance comprehension of the Arabic QA research environment, this study presents a taxonomy of current systems categorized by technique, domain scope, input-output formats, and DL architecture. This is succeeded by an examination of existing research deficiencies and unresolved difficulties.

A. Fatwa QA Dataset (In Progress)

The Fatwa QA dataset under development aims to fill the significant gap in Arabic QA datasets that are specific to certain fields, such as religion and law. To ensure that the dataset is real and important for academics, it is being put together from verified online Fatwa archives, which cover a wide scope of Islamic law, such as family law, worship, and business. The final version of the dataset should have not only factual questions but also opinion-based questions that require understanding the context and thinking about it. The MSA will be used to write each QA pair, with standard vocabulary and direct quotes from reliable sources such as the Qur'an, Hadith, and classical legal studies. The dataset is made to work with both extractive and abstractive QA tasks, which allows a wide range of neural modeling methods to be used. This resource will be necessary for future experimental studies on Arabic QA and will serve as a standard for assessing domain-specific models. The dataset is still being worked on, but it will have a lot of questions and answers about Fatwa, making it possible to systematically investigate religious domain QA tasks.

B. Taxonomy of Arabic QA Systems

Arabic QA systems may be classified across several aspects, each representing unique design objectives and functionalities. Initially, systems may be categorized according to domain scope as either closed-domain, specifically designed for particular fields such as education, religion, or law (e.g., ARCD) [17], or open-domain, developed using general-purpose corpora like Wikipedia, exemplified by datasets such as Arabic-SQuAD and TyDiQA-Arabic [17, 66].

Regarding model type, Arabic QA methods encompass extractive models that pinpoint text spans within source documents (e.g., AraBERT fine-tuned on Arabic-SQuAD) [20], abstractive models that produce answers in natural language (e.g., ArabicT5, mBART) [22, 90], and hybrid models that amalgamate dense retrieval with neural generation (e.g., RAG-based systems) [29]. Moreover, systems vary according to input modality [22, 90]. Although most existing Arabic QA systems are text-centric, there is a burgeoning interest in multimodal QA that integrates images, audio, or OCR-processed documents [102, 103], yet practical applications in Arabic remain limited [65]. Ultimately, another essential dimension of categorization is linguistic diversity. Most systems function on MSA, which predominates in academic corpora and benchmarks [20]. However, dialectal Arabic, prevalent in various locations, is still inadequately represented, despite the backing of models such as MARBERT [21] and minimal initiatives to develop dialectal QA datasets [41]. These categorization dimensions provide a systematic framework to examine the progression and variety of Arabic QA systems, pinpointing neglected areas for future advances.

TABLE V. TAXONOMY OF ARABIC QA SYSTEMS

Criterion	Category	Examples / Notes
Domain scope	Closed-domain	Legal, religious, and educational QA (e.g., ARCD)
	Open-domain	Wikipedia-based QA (e.g., Arabic-SQuAD, TyDiQA-Arabic)
Model type	Extractive	Span-based answers (e.g., AraBERT on Arabic-SQuAD)
	Abstractive	Natural language generation (e.g., ArabicT5, mBART)
	Hybrid (RAG)	RAG
Language variety	MSA	MSA dominates current research and datasets
	Dialectal	Informal/regional Arabic (e.g., MARBERT, MADAR corpus)
Input modality	Text-only	Most Arabic QA systems are currently text-based
	Multimodal (future)	Integrating text with image/audio (e.g., OCR-based Arabic literature QA)
Answer format	Extractive span	Retrieved directly from context
	Generated text	Synthesized responses via Seq2Seq models

C. Research Gaps and Open Challenges

Despite the considerable advances in Arabic QA, several essential deficiencies persist, obstructing the evolution of practical and resilient systems. A significant constraint is the paucity of domain-specific datasets, as most publicly accessible Arabic QA datasets, such as Arabic-SQuAD and ARCD, are sourced from Wikipedia and exhibit insufficient representation in specialized fields [17], [63]. This limits the capacity of QA systems to generalize across many use cases. The field is predominantly driven by extractive paradigms, with limited exploration of generative techniques. There is a significant deficiency of high-quality abstractive datasets and extensive implementations of generative models [22]. In addition, Arabic dialects—although widespread in social networks, consumer relations, and verbal communication—are significantly underrepresented, as most systems are trained on MSA, limiting their practical applicability [40].

Another issue is the variability in assessment methods. Research frequently varies in the datasets and metrics employed, with numerous studies neglecting human evaluation, leading to disjointed benchmarking and challenges in equitable model comparison [104]. Furthermore, promising hybrid architectures, such as RAG, which have demonstrated significant efficacy in English QA, are infrequently utilized in Arabic, despite their benefits in data-scarce and knowledge-intensive applications [29]. Finally, explainability and trustworthiness are inadequately addressed, especially in sensitive areas such as religious fatwas, where users seek not only accurate responses but also transparent justifications and reliable sources—a challenge that opaque DL models frequently do not fulfill [105]. Rectifying these deficiencies is essential for enhancing Arabic QA, rendering it more inclusive, reliable, and contextually informed.

IX. CHALLENGES AND FUTURE DIRECTIONS

A. General Challenges in Arabic QA

Arabic QA research is growing; however, certain challenges persist, offering significant prospects for innovation and wider influence. A viable future avenue is the creation of domain-specific QA solutions. Current Arabic QA systems are mostly developed using general-purpose corpora, although there is an increasing demand for applications customized for specific areas such as religion, healthcare, education, and law. Al-Bayan, an Arabic QA system designed for the Holy Quran, illustrates the potential for specialized applications [106].

Future studies should focus on improving generative QA. Most existing systems are predominantly extractive; however, utilizing advanced generative designs such as AraT5 and mBART can be used to develop coherent and semantically precise replies. The implementation of controlled generation techniques and the integration of source citation procedures are essential to increase user confidence, particularly in sensitive areas such as medical care or religious fatwas [107]. Dialectal and code-switched QA constitutes another significant deficiency. Considering the extensive variety of Arabic dialects and their prominence in practical communication, future systems should incorporate dialectal models such as MARBERT [21] and develop QA corpora that represent these variations. In addition, Arabic QA research might benefit from exploring multimodal domains. Current initiatives primarily focus on text, and the incorporation of diverse data formats—including images, audio, or OCR-processed documents—might provide novel functionalities, such as QA regarding scanned religious rulings or verbal fatwas [108].

The use of RAG in Arabic QA is limited, despite its efficacy in English-language applications [29]. Due to the limited resources available for Arabic, hybrid systems that separate retrieval from creation provide a scalable approach. Optimizing retrievers in legal corpora may facilitate high recall, evidence-based creation. A prospective avenue entails few-shot and cross-lingual learning, particularly given the increased annotation costs for Arabic. Research on mT5 and XLM-R indicates that zero-shot and few-shot transfer from English can provide favorable outcomes for low-resource Arabic QA tasks [62].

Ultimately, human-centered assessment and trust are crucial in delicate fields such as law and religion. Future Arabic QA systems must go beyond accuracy alone, prioritizing explainability, transparency, and user satisfaction, especially through culturally relevant and interpretable outputs [108]. The study in [109] looked at hybrid architectures that integrate LLMs with structured ontologies. These structures might improve semantic control and domain alignment, which are important for scaling QA in domains such as religion or constitutional Arabic. Implementing these progressive tactics will advance Arabic QA research toward practical applicability and readiness.

B. Recent Advances in Arabic RAG Systems

Recent research has begun the adaptation of RAG frameworks to Arabic, addressing previous constraints of monolingual generative or extractive models. ARAG [110] is an RAG-based architecture that incorporates semantic embeddings specifically designed for Arabic, enhancing document retrieval and generative coherence. In [111], several embedding-based retrieval methods were evaluated in Arabic lexical QA, revealing that DPR-based retrievers paired with Arabic-optimized generators (such as ArabicT5 or mBART) enhance both response fluency and factual accuracy. Similarly, in [59], optimization methods for Arabic RAG pipelines were examined, encompassing Facebook AI Similarity Search (FAISS) index tweaking and passage reranking, and underscored the advantages of integrating multilingual generators with domain-specific Arabic indices. These contributions highlight the potential of RAG for Arabic QA, especially in low-resource or domain-specific scenarios similar to religious questions. However, deficiencies persist in expanding these methods to encompass a greater variety of dialects, including multimodal data, and assessing generative fidelity. The insights of these studies into future system design will be essential for the development of resilient, interpretable, and contextually relevant Arabic QA systems.

C. Future Directions for Arabic RAG

Implementing RAG in Arabic QA entails potential and challenges. The morphological complexity and dialectal variety of Arabic hinder dense retrieval and tokenization [21], [112]. The utilization of multilingual or Arabic-specific retrievers, such as MARBERT [21] and QARiB [113], in conjunction with particular Arabic document indices, can substantially improve the performance of RAG [29]. Furthermore, RAG's separation of retrieval and production facilitates modular fine-tuning, allowing researchers to modify the retriever for dialectal or domain-specific corpora while maintaining a multilingual generator such as mBART or fine-tuning it on Arabic QA datasets [29].

RAG is a scalable framework for enhancing research in Arabic QA, providing several interesting future avenues. A significant milestone is the creation of dense retrievers pretrained on Arabic text using contrastive learning objectives, facilitating enhanced document retrieval in morphologically intricate contexts [33]. Simultaneously, the development of extensive and varied Arabic corpora tailored for open-domain QA will be essential to improve the coverage and accuracy of the retrieval component [40]. Furthermore, researchers can

investigate the zero- and few-shot capabilities of RAG by utilizing multilingual prompts, which might mitigate data scarcity in Arabic-specific QA tasks [6]. A further compelling approach is the integration of RAG with dialect identification methods, facilitating the system's ability to comprehend and respond to mixed language or dialectal inquiries with more precision [21].

As a result, RAG signifies a pivotal advancement in the development of QA systems. Despite its preliminary application to Arabic QA, the framework presents substantial promise for innovative contributions in both retrieval and generative modeling, particularly in low-resource and linguistically varied environments [29].

X. CONCLUSION

With an emphasis on DL techniques, this review examined the state of Arabic QA systems today. Although there has been much progress recently, this analysis reveals lingering issues, such as the dominance of extractive methods, the lack of RAG integration, and the inadequate adaptation of multilingual models for the particular linguistic complexity of Arabic. This study highlights the need for increased investment in domain-specific datasets, better management of dialectal and morphologically rich forms of Arabic, and improved evaluation frameworks that take faithfulness and explainability into account, along with accuracy, to close these gaps.

This research suggests a domain-specific Arabic QA system based on Fatwas for future work, backed by a cutting-edge dataset that is currently being developed. This dataset, compiled from verified religious portals in Saudi Arabia and throughout the MENA region, is intended to enable retrieval-augmented generative QA capable of producing precise and well-founded answers. This resource offers a suggested path for developing domain-specific Arabic QA research rather than as a finished product.

In conclusion, Arabic QA is still a relatively unexplored but exciting field in NLP. This work helps the community develop a more defined research agenda by synthesizing previous studies and defining future directions, especially in domain-specific applications such as Fatwas.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 2025.

[2] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *Natural Language Engineering*, vol. 7, no. 4, pp. 275–300, Dec. 2001, https://doi.org/10.1017/S1351324901002807.

[3] E. M. Voorhees and D. M. Tice, "The TREC-8 Question Answering Track," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, Feb. 2000.

[4] J. H. Martin and D. Jurafsky, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, vol. 23. 2009.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, Mar. 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[6] T. Brown *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[7] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, 2020.

[8] X. Li and D. Roth, "Learning Question Classifiers," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[9] E. M. Voorhees and D. M. Tice, "The TREC-8 Question Answering Track Evaluation," *NIST*, vol. 3, May 2000.

[10] S. Badugu and R. Manivannan, "A study on different closed domain question answering approaches," *International Journal of Speech Technology*, vol. 23, no. 2, pp. 315–325, Jun. 2020, https://doi.org/10.1007/s10772-020-09692-0.

[11] A. Soudi, G. Neumann, and A. van den Bosch, "Arabic Computational Morphology: Knowledge-based and Empirical Methods," in *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, A. Soudi, A. van den Bosch, and G. Neumann, Eds. Springer Netherlands, 2007, pp. 3–14.

[12] Z. Abbasiantaeb and S. Momtazi, "Text-based Question Answering from Information Retrieval and Deep Neural Network Perspectives: A Survey." arXiv, May 27, 2020, https://doi.org/10.48550/arXiv.2002.06612.

[13] H. A. Pandya and B. S. Bhatt, "Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices." arXiv, Dec. 07, 2021, https://doi.org/10.48550/arXiv.2112.03572.

[14] E. M. Bender, "On achieving and evaluating language-independence in NLP," *Linguistic Issues in Language Technology*, vol. 6, 2011.

[15] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Apr. 2020, pp. 6282–6293, https://doi.org/10.18653/v1/2020.acl-main.560.

[16] N. Y. Habash, *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers, 2010.

[17] H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic Question Answering," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Florence, Italy, Dec. 2019, pp. 108–118, https://doi.org/10.18653/v1/W19-4612.

[18] R. Malhas, W. Mansour, and T. Elsayed, "Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an," 2023, https://doi.org/10.18653/v1/2023.arabicnlp-1.76.

[19] R. Malhas, W. Mansour, and T. Elsayed, "Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an," in *Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, 2023, pp. 690–701, https://doi.org/10.18653/v1/2023.arabicnlp-1.76.

[20] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding." arXiv, Mar. 07, 2021, https://doi.org/10.48550/arXiv.2003.00104.

[21] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations." arXiv, Feb. 21, 2021, https://doi.org/10.48550/arXiv.2102.10684.

[22] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-Text Transformers for Arabic Language Generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, Feb. 2022, pp. 628–647, https://doi.org/10.18653/v1/2022.acl-long.47.

[23] O. Kolomiyets and M. F. Moens, "A survey on question answering technology from an information retrieval perspective," *Information Sciences*, vol. 181, no. 24, pp. 5412–5434, Dec. 2011, https://doi.org/10.1016/j.ins.2011.07.047.

[24] A. Rogers, M. Gardner, and I. Augenstein, "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension," *ACM Computing Surveys*, vol. 55, no. 10, Oct. 2023, Art. no. 197, https://doi.org/10.1145/3560260.

[25] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 2383–2392, https://doi.org/10.18653/v1/D16-1264.

[26] E. Dai *et al.*, "A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability," *Machine Intelligence Research*, vol. 21, no. 6, pp. 1011–1061, Dec. 2024, https://doi.org/10.1007/s11633-024-1510-8.

[27] Z. Yang *et al.*, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct. 2018, pp. 2369–2380, https://doi.org/10.18653/v1/D18-1259.

[28] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing Datasets for Multi-hop Reading Comprehension Across Documents," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 287–302, May 2018, https://doi.org/10.1162/tacl_a_00021.

[29] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv, Apr. 12, 2021, https://doi.org/10.48550/arXiv.2005.11401.

[30] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions." arXiv, Apr. 28, 2017, https://doi.org/10.48550/arXiv.1704.00051.

[31] D. Ferrucci *et al.*, "Building Watson: An Overview of the DeepQA Project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, Jul. 2010, https://doi.org/10.1609/aimag.v31i3.2303.

[32] D. Moldovan *et al.*, "The structure and performance of an open-domain question answering system," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, USA, Jul. 2000, pp. 563–570, https://doi.org/10.3115/1075218.1075289.

[33] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Aug. 2020, pp. 6769–6781, https://doi.org/10.18653/v1/2020.emnlp-main.550.

[34] M. Artetxe and H. Schwenk, "Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Apr. 2019, pp. 3197–3203, https://doi.org/10.18653/v1/P19-1309.

[35] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, PA, USA, 2001, https://doi.org/10.3115/1073083.1073135.

[36] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Apr. 2004, pp. 74–81.

[37] C. W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 2122–2132, https://doi.org/10.18653/v1/D16-1230.

[38] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "Evaluating Question Answering Evaluation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Hong Kong, China, Aug. 2019, pp. 119–124, https://doi.org/10.18653/v1/D19-5817.

[39] E. Durmus, H. He, and M. Diab, "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Apr. 2020, pp. 5055–5070, https://doi.org/10.18653/v1/2020.acl-main.454.

[40] S. Mazzucchi, N. Leone, S. Azzini, L. Pavesi, and V. Moretti, "Entropy certification of a realistic quantum random-number generator based on single-particle entanglement," *Physical Review A*, vol. 104, no. 2, Aug. 2021, Art. no. 022416, https://doi.org/10.1103/PhysRevA.104.022416.

[41] H. Bouamor et al., "The MADAR Arabic dialect corpus and lexicon," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[42] A. Rozovskaya, H. Bouamor, N. Habash, W. Zaghouani, O. Obeid, and B. Mohit, "The Second QALB Shared Task on Automatic Text Correction for Arabic," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, 2015, pp. 26–35, https://doi.org/10.18653/v1/W15-3204.

[43] M. Nabil et al., "AlQuAnS – An Arabic Language Question Answering System:," in Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Madeira, Portugal, 2017, pp. 144–154, https://doi.org/10.5220/0006602901440154.

[44] L. Abouenour, K. Bouzoubaa, and P. Rosso, "On the extension of arabic wordnet named entities and its impact on question / answering," presented at the International Conference on Knowledge Engineering and Ontology Development, Oct. 2010, vol. 2, pp. 424–429, https://doi.org/10.5220/0003102004240429.

[45] E. Noguera, F. Llopis, and A. Ferrández, "Passage Filtering for Open-Domain Question Answering," in *Advances in Natural Language Processing*, 2006, pp. 534–540, https://doi.org/10.1007/11816508_53.

[46] 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, Jan. 2022, https://doi.org/10.1007/s10462-021-10031-1.

[47] Y. Alkhurayyif and A. R. W. Sait, "A comprehensive survey of techniques for developing an Arabic question answering system," *PeerJ Computer Science*, vol. 9, Jun. 2023, Art. no. e1413, https://doi.org/10.7717/peerj-cs.1413.

[48] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," *Procedia Computer Science*, vol. 73, pp. 366–375, Jan. 2015, https://doi.org/10.1016/j.procs.2015.12.005.

[49] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no. 3, pp. 345–361, Jul. 2016, https://doi.org/10.1016/j.jksuci.2014.10.007.

[50] M. Essam, M. A. Deif, and R. Elgohary, "Deciphering Arabic question: a dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways," *Artificial Intelligence Review*, vol. 57, no. 9, Aug. 2024, Art. no. 251, https://doi.org/10.1007/s10462-024-10880-6.

[51] H. M. Al Chalabi, S. K. Ray, and K. Shaalan, "Question classification for Arabic Question Answering Systems," in *2015 International Conference on Information and Communication Technology Research (ICTRC)*, Feb. 2015, pp. 310–313, https://doi.org/10.1109/ICTRC.2015.7156484.

[52] W. Yih, M. W. Chang, C. Meek, and A. Pastusiak, "Question Answering Using Enhanced Lexical Semantic Models," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria, Dec. 2013, pp. 1744–1753.

[53] F. T. AL-Khawaldeh, "Answer Extraction for Why Arabic Questions Answering Systems: EWAQ." arXiv, Jul. 04, 2019, https://doi.org/10.48550/arXiv.1907.04149.

[54] A. Pasha *et al.*, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, Feb. 2014, pp. 1094–1101.

[55] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, https://doi.org/10.1126/science.aaa8685.

[56] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, Dec. 2018, https://doi.org/10.1109/MCI.2018.2840738.

[57] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining." arXiv, Jan. 22, 2019, https://doi.org/10.48550/arXiv.1901.07291.

[58] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." arXiv, Oct. 29, 2019, https://doi.org/10.48550/arXiv.1910.13461.

[59] S. R. El-Beltagy and M. A. Abdallah, "Exploring Retrieval Augmented Generation in Arabic." arXiv, Aug. 14, 2024, https://doi.org/10.48550/arXiv.2408.07425.

[60] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-Shot Cross-Lingual Transfer with Meta Learning." arXiv, Oct. 05, 2020, https://doi.org/10.48550/arXiv.2003.02739.

[61] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale." arXiv, Apr. 08, 2020, https://doi.org/10.48550/arXiv.1911.02116.

[62] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv, Mar. 11, 2021, https://doi.org/10.48550/arXiv.2010.11934.

[63] A. Abdallah *et al.*, "ArabicaQA: A Comprehensive Dataset for Arabic Question Answering." arXiv, Mar. 26, 2024, https://doi.org/10.48550/arXiv.2403.17848.

[64] M. Attia and A. Elkahky, "Segmentation for Domain Adaptation in Arabic," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Florence, Italy, Dec. 2019, pp. 119–129, https://doi.org/10.18653/v1/W19-4613.

[65] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, Aug. 2021, pp. 6974–6996, https://doi.org/10.18653/v1/2021.emnlp-main.559.

[66] A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual Open-Retrieval Question Answering." arXiv, Apr. 13, 2021, https://doi.org/10.48550/arXiv.2010.11856.

[67] M. A. Daoud, C. Abouzahir, L. Kharouf, W. Al-Eisawi, N. Habash, and F. E. Shamout, "MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks." arXiv, Aug. 22, 2025, https://doi.org/10.48550/arXiv.2505.03427.

[68] M. AL-Qurishi, S. AlQaseemi, and R. Soussi, "AraLegal-BERT: A pretrained language model for Arabic Legal text." arXiv, Oct. 15, 2022, https://doi.org/10.48550/arXiv.2210.08284.

[69] R. Mohammad, O. S. Alkhnbashi, and M. Hammoudeh, "Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications," *Big Data and Cognitive Computing*, vol. 8, no. 11, Nov. 2024, Art. no. 157, https://doi.org/10.3390/bdcc8110157.

[70] E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas, "A survey on question answering systems over linked data and documents," *Journal of Intelligent Information Systems*, vol. 55, no. 2, pp. 233–259, Oct. 2020, https://doi.org/10.1007/s10844-019-00584-7.

[71] M. Sheker, S. Saad, R. Abood, and M. Shakir, "Domain-specific ontology-based approach for Arabic question answering," *Journal of Theoretical and Applied Information Technology*, vol. 83, no. 1, pp. 43–51, 2016.

[72] K. Benlaharche, Z. Laboudi, N. Nouaouria, and D. E. Zegour, "An ontology driven question answering system for fatawa retrieval," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 980–992, Aug. 2021, https://doi.org/10.11591/ijeecs.v23.i2.pp980-992.

[73] Z. Saadaoui, G. Tlig, and F. Jarray, "LLMs Based Approach for Quranic Question Answering:," in *Proceedings of the 20th International Conference on Web Information Systems and Technologies*, Porto, Portugal, 2024, pp. 112–118, https://doi.org/10.5220/0013012900003825.

[74] F. Qamar, S. Latif, and R. Latif, "A Benchmark Dataset with Larger Context for Non-Factoid Question Answering over Islamic Text." arXiv, Sep. 15, 2024, https://doi.org/10.48550/arXiv.2409.09844.

[75] Z. Khalila *et al.*, "Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, 2025, https://doi.org/10.14569/IJACSA.2025.01602134.

[76] A. Mostafa and O. Mohamed, "GOF at Qur'an QA 2022: Towards an Efficient Question Answering For The Holy Qu'ran In The Arabic Language Using Deep Learning-Based Approach," in *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, Marseille, France, Mar. 2022, pp. 104–111.

[77] S. Maged *et al.*, "HistoryQuest: Arabic Question Answering in Egyptian History with LLM Fine-Tuning and Transformer Models," in *2024 Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt, Jul. 2024, pp. 135–140, https://doi.org/10.1109/IMSA61967.2024.10652824.

[78] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora." arXiv, Feb. 25, 2017, https://doi.org/10.48550/arXiv.1702.07835.

[79] K. Darwish, H. Mubarak, and A. Abdelali, "Arabic Diacritization: Stats, Rules, and Hacks," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, Valencia, Spain, Dec. 2017, pp. 9–17, https://doi.org/10.18653/v1/W17-1302.

[80] K. Darwish, H. Sajjad, and H. Mubarak, "Verifiably Effective Arabic Dialect Identification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Jul. 2014, pp. 1465–1468, https://doi.org/10.3115/v1/D14-1154.

[81] K. Darwish, "Arabizi Detection and Conversion to Arabic." arXiv, Jun. 28, 2013, https://doi.org/10.48550/arXiv.1306.6755.

[82] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological Analysis and Disambiguation for Dialectal Arabic," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, USA, Mar. 2013, pp. 426–432.

[83] S. Hossain, F. Shammary, B. Shammary, and H. Afli, "Enhancing Dialectal Arabic Intent Detection through Cross-Dialect Multilingual Input Augmentation," in *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, Abu Dhabi, United Arab Emirates, Jan. 2025, pp. 44–49.

[84] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, "QADI: Arabic Dialect Identification in the Wild," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine Dec. 2021.

[85] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA, USA, Mar. 2016, pp. 11–16, https://doi.org/10.18653/v1/N16-3003.

[86] Y. Belinkov and J. Glass, "Arabic Diacritization with Recurrent Neural Networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Jun. 2015, pp. 2281–2285, https://doi.org/10.18653/v1/D15-1274.

[87] A. H. Alshehri, "AraQA-BERT: Towards an Arabic Question Answering System using Pre-trained BERT Models," *WSEAS Transactions on Information Science and Applications*, vol. 21, pp. 361–373, 2024, https://doi.org/10.37394/23209.2024.21.34.

[88] M. A. Ali, N. Daftardar, M. Waheed, J. Qin, and D. Wang, "MQA-KEAL: Multi-hop Question Answering under Knowledge Editing for Arabic Language." arXiv, Sep. 18, 2024, https://doi.org/10.48550/arXiv.2409.12257.

[89] R. Tsarfaty, D. Bareket, S. Klein, and A. Seker, "From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)?" arXiv, May 04, 2020, https://doi.org/10.48550/arXiv.2005.01330.

[90] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, Nov. 2020, https://doi.org/10.1162/tacl_a_00343.

[91] F. Petroni *et al.*, "KILT: a Benchmark for Knowledge Intensive Language Tasks." arXiv, May 27, 2021, https://doi.org/10.48550/arXiv.2009.02252.

[92] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, "Deep Learning for Answer Sentence Selection." arXiv, Dec. 04, 2014, https://doi.org/10.48550/arXiv.1412.1632.

[93] K. M. Hermann *et al.*, "Teaching Machines to Read and Comprehend." arXiv, Nov. 19, 2015, https://doi.org/10.48550/arXiv.1506.03340.

[94] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Aug. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[95] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 02, 2023, https://doi.org/10.48550/arXiv.1706.03762.

[96] Y. Tang *et al.*, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning." arXiv, Aug. 02, 2020, https://doi.org/10.48550/arXiv.2008.00401.

[97] E. Chang, A. Marin, and V. Demberg, "Programmable Annotation with Diversed Heuristics and Data Denoising," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, Jul. 2022, pp. 2681–2691.

[98] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, Dubai, United Arab Emirates, Nov. 2024, pp. 607–611, https://doi.org/10.1109/FLLM63129.2024.10852457.

[99] M. Alshammary, M. N. Uddin, and L. Khan, "RFPG: Question-Answering from Low-Resource Language (Arabic) Texts using Factually Aware RAG," in *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, Washington, DC, USA, Oct. 2024, pp. 107–116, https://doi.org/10.1109/CIC62241.2024.00023.

[100] M. Alsuhaibani and M. O. Beg, "Improving Domain-Specific Data Question Answering with Deep and Cross-Lingual Transfer Learning," in *Machine Learning and Soft Computing*, 2025, pp. 80–93, https://doi.org/10.1007/978-981-96-6400-9_7.

[101] R. Malhas, W. Mansour, and T. Elsayed, "Qur'an QA 2022: Overview of The First Shared Task on Question Answering over the Holy Qur'an," in *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, Marseille, France, Mar. 2022, pp. 79–87.

[102] X. Yuan *et al.*, "Multimodal Contrastive Training for Visual Representation Learning." arXiv, Apr. 26, 2021, https://doi.org/10.48550/arXiv.2104.12836.

[103] J. B. Alayrac *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning." arXiv, Nov. 15, 2022, https://doi.org/10.48550/arXiv.2204.14198.

[104] A. Hegde, A. Kumar, A. Agarwala, and B. Muralidharan, "Exploring ideas in topological quantum phenomena: A journey through the SSH model." arXiv, Aug. 03, 2021, https://doi.org/10.48550/arXiv.2108.01460.

[105] S. Alnefaie, E. Atwell, and M. A. Alsalka, "Qur'an Passage Ranking Using Transformer Models," in *Arabic Language Processing: From Theory to Practice*, 2025, pp. 183–194, https://doi.org/10.1007/978-3-031-79164-2_16.

[106] H. Abdelnasser *et al.*, "Al-Bayan: An Arabic Question Answering System for the Holy Quran," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar, Jul. 2014, pp. 57–64, https://doi.org/10.3115/v1/W14-3607.

[107] A. Ghaddar *et al.*, "Revisiting Pre-trained Language Models and their Evaluation for Arabic Natural Language Understanding." arXiv, May 21, 2022, https://doi.org/10.48550/arXiv.2205.10687.

[108] A. Alrayzah, F. Alsolami, and M. Saleh, "Challenges and opportunities for Arabic question-answering systems: current techniques and future directions," *PeerJ Computer Science*, vol. 9, Oct. 2023, Art. no. e1633, https://doi.org/10.7717/peerj-cs.1633.

[109] S. Alamoudi, L. A. A. Khuzayem, and A. Jamal, "Optimizing Automated Question Generation for Educational Assessments: A Semantic Analysis of LLMs with Structured and Unstructured Ontologies," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23664–23671, Jun. 2025, https://doi.org/10.48084/etasr.10662.

[110] H. Abdelazim, M. Tharwat, and A. Mohamed, "Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, https://doi.org/10.14569/IJACSA.2023.01411135.

[111] R. Al-Rasheed *et al.*, "Evaluating RAG Pipelines for Arabic Lexical Information Retrieval: A Comparative Study of Embedding and Generation Models," in *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, Abu Dhabi, UAE, Jan. 2025, pp. 155–164.

[112] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," *Applied Sciences*, vol. 12, no. 11, Jan. 2022, Art. no. 5720, https://doi.org/10.3390/app12115720.

[113] H. Mulki and B. Ghanem, "Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual), Dec. 2021, pp. 154–163.