Machine learning and deep learning techniques in Arabic question answering systems: innovations and challenges

Azza Mohamed[1], Khaled Abdelqader[2] and Khaled Shaalan[2]

[1] College of Engineering and Computing, Liwa University, Al Ain, United Arab Emirates
[2] Faculty of Engineering and IT, British University in Dubai, Dubai, United Arab Emirates

Submitted 12 August 2024
Accepted 6 October 2025
Published 28 November 2025

Corresponding author
Azza Mohamed,
azza.basiouni@lu.ac.ae

Academic editor
Xiangjie Kong

Additional Information and Declarations can be found on page 32

DOI 10.7717/peerj-cs.3331

ABSTRACT

Research on Arabic question answering improves information access, promotes language variety, advances Arabic Language Processing NLP technology, and has educational, cultural, economic, and societal implications. This study delves deeply into Artificial Intelligence (AI)-based Arabic Question Answering Systems (AQAS), with an emphasis on the use of machine learning (ML) and deep learning (DL) technologies to improve Arabic language processing and comprehension. A careful analysis of twelve qualifying studies done between 2018 and 2023 identifies considerable advances in the use of advanced computational approaches to address Arabic's distinctive linguistic problems. This work is particularly relevant for practitioners and researchers in the fields of AI and NLP, as well as professionals interested in the implications of AI technologies for Arabic language processing. Moreover, we recognize that the impact of AQAS extends beyond academia; it has significant implications for various sectors, including education, technology, and information access. Through this comprehensive examination, we aim to lay the groundwork for ongoing innovation and development in AQAS. Our study emphasizes the necessity for high-quality, diverse datasets that encompass the linguistic variety and dialectal variations of Arabic. We also explore the potential of hybrid AI models for improved semantic analysis, while acknowledging the computational challenges faced by state-of-the-art AI models. The study also acknowledges the computational challenges faced by state-of-the-art AI models and suggests future research directions focused on developing lightweight, efficient models, enhancing semantic analysis, and ensuring the fairness and equity of AI applications. Despite substantial progress, the study identifies gaps in handling linguistic nuances, the scarcity of annotated datasets, and the limited exploration of innovative AI techniques. It calls for a collaborative effort to enrich the Arabic Question Answering (AQA) datasets, improve computational efficiency and advance the semantic understanding capabilities of AQAS.

Subjects Bioinformatics, Artificial Intelligence, Data Mining and Machine Learning, Natural Language and Speech
Keywords Machine learning, Deep learning, AI, NLP, Systematic review, Question answering systems

INTRODUCTION

Question answering (QA) systems demonstrate artificial intelligences (AIs) progress in mimicking human comprehension and response (*Hicke et al., 2023*). These systems use

How to cite this article Mohamed A, Abdelqader K, Shaalan K. 2025. Machine learning and deep learning techniques in Arabic question answering systems: innovations and challenges. **PeerJ Comput. Sci. 11:e3331 DOI 10.7717/peerj-cs.3331**

natural language processing (NLP), machine learning (ML), and knowledge management to parse and react to user questions (*Sarkar, Gupta & Singh, 2023*). QA systems are classified as factoid or non-factoid, open-domain, or closed domain, and each caters to a distinct sort of query. Early QA systems used rule-based methods (*Mervin, 2013*), but newer systems incorporate deep learning and transfer learning for more complex context comprehension (*Ishwari et al., 2019*). Knowledge graphs and databases complement these systems (*Sequeda, Allemang & Jacob, 2023*). Language complexity and the necessity for domain-specific knowledge remain significant challenges (*El Adlouni et al., 2021*; *Lahbari, El Alaoui & Zidani, 2018*; *Li, 2018*; *Sultana & Badugu, 2020*). However, AI has made QA systems more accessible and capable of automating knowledge acquisition; hence, increasing human-machine interaction (*Wang et al., 2022*).

Integration of components in question answering systems

A typical AQAS has a hierarchical architecture that includes three main components: Question Analysis, Passage Retrieval, and Response Extraction (Fig. 1). Preprocessing (tokenization, normalization, named entity recognition) and query expansion are used in the Question Analysis stage to classify question types. The Passage Retrieval step selects and rates candidate texts from huge corpora using retrieval models like BM25 or neural methods. Finally, the Response Extraction stage employs answer selection and validation algorithms to generate the most accurate result. These stages combine to build an integrated pipeline that solves Arabic's linguistic complexity.

Figure 1 depicts the general design of an Arabic Question Answering System (AQAS), including the three main stages: (1) Question Analysis-preprocessing, classification, and input expansion; (2) Passage Retrieval-identifying and ranking relevant text segments; and (3) Response Extraction-selecting and validating the final response. This diagram shows how the components interact to manage the linguistic complexity of Arabic.

The first phase, question analysis, requires multiple preprocessing techniques such as tokenization, normalization, and named entity identification. This phase also includes query expansion techniques like synonym substitution and question classification to determine the type of response needed (*Malik, Sharan & Biswas, 2013*).

Following question analysis, the system enters the information retrieval phase. This process begins with document collecting and progresses to passage retrieval, where relevant texts are filtered according to the user's query. In this phase, candidate responses are extracted and ranked according to their relevance to the topic given (*Bhoir & Potey, 2014*; *Calijorne Soares & Parreiras, 2020*).

Finally, during the response extraction step, the system selects the best feasible answer applying specialized algorithms. This process may also involve answer validation, which examines the confidence levels of the selected responses (*Al Chalabi, 2015*; *Alwaneen et al., 2022*).

By weaving these elements into a cohesive story, we demonstrate the complex procedures involved in designing efficient question answering systems, particularly when dealing with the specific constraints given by the Arabic language.

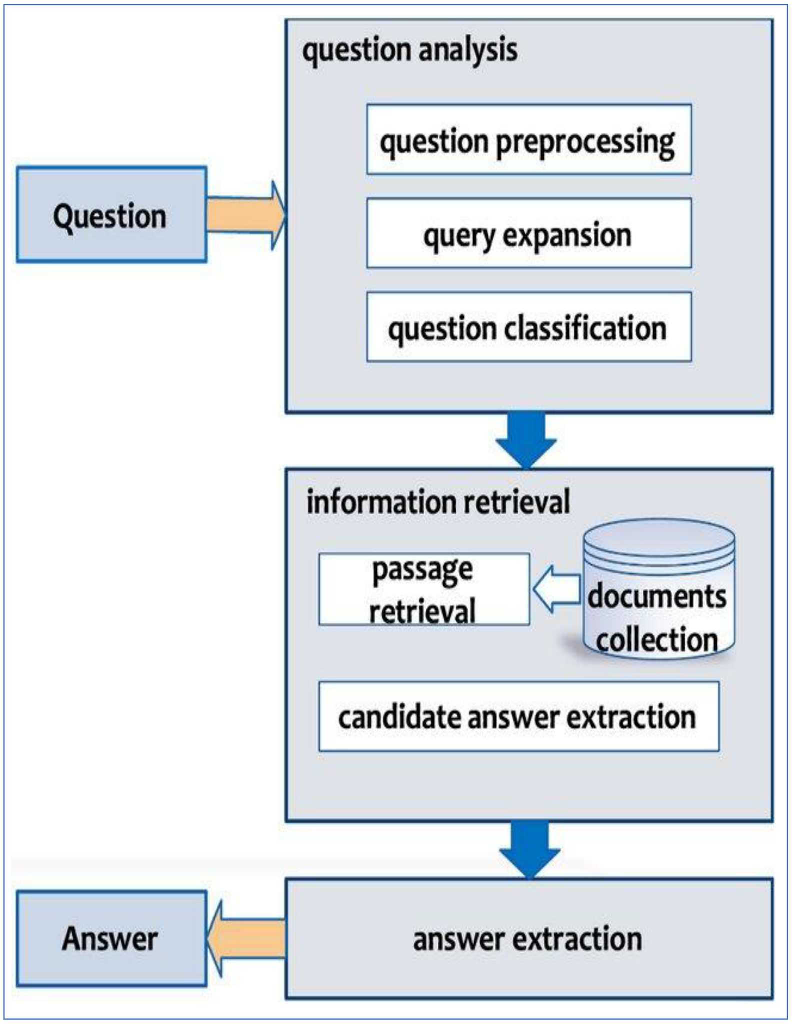

Figure 1 **QAS high level diagram (*Alwaneen et al., 2022*).** Full-size DOI: 10.7717/peerj-cs.3331/fig-1

Significance and challenges of Arabic question answering systems

Significant research work has advanced QAS for English and other European languages. However, the development of Arabic QAS has been slowed due to the language's inherent problems as well as a scarcity of specialized NLP tools and datasets (*Al Chalabi, 2015*; *Abdelqader, Mohamed & Shaalan, 2023*). Arabic is one of the world's most spoken languages, with over 400 million speakers in more than 25 nations. It has a tremendous cultural, educational, and technological impact. Furthermore, Arabic is one of the United Nations' six official languages, demonstrating its global importance and relevance in international communication and policymaking (*UNESCO, 2021*; *United Nations, 2020*). Between 2000 and 2015, the language saw substantial growth on the internet. Despite this, Arabic poses distinct hurdles because of its rich morphology and structural intricacies. Language processing can be challenging to automate due to features such as sophisticated word-building and context-dependent interpretations. Additionally, the frequent removal of diacritical marks from texts significantly complicates NLP operations (*Al-Saleh & Menai, 2016*; *El-Deeb, Al-Zoghby & Elmougy, 2018*). Additional issues include morphological modifications that depend on letter location. There is also a lack of capitalization to differentiate nouns. Furthermore, the variety of word forms caused by roots, patterns, and affixes complicates processing (*El-Deeb, Al-Zoghby & Elmougy, 2018*; *Zaki, 2019*). These linguistic traits present considerable challenges for both Arab and non-Arab scholars. They complicate the design of successful NLP tools and linguistic resources, particularly for AQAS. This situation indicates a significant opportunity for future research (*Elsaid et al., 2022*).

Overview of Arabic NLP development and challenges

The field of NLP has advanced significantly over the years, driven by the increasing demand for improved Arabic information retrieval (*Alkhurayyif & Sait, 2023*; *Alwaneen et al., 2022*). A structured staging of NLP progress is necessary to highlight key historical milestones, advancements, and trends, particularly in AQAS.

Historical progress of Arabic NLP

- Early rule-based systems (Pre-2000s): Initial efforts in Arabic NLP relied on handcrafted linguistic rules for tasks such as morphological analysis and syntax parsing. However, these systems struggled with scalability and robustness.
- Statistical NLP and machine learning (2000s–2010s): The introduction of probabilistic models and feature-based machine learning significantly improved Arabic text processing. Named Entity Recognition (NER) and Part-of-Speech (POS) tagging saw notable advancements.
- Deep learning and Transformer-based models (2018–Present): The emergence of pre-trained models like AraBERT and multilingual transformers revolutionized Arabic NLP, particularly in AQAS, enabling context-aware processing and improved syntactic parsing (*El Adlouni et al., 2021*; *Lahbari, El Alaoui & Zidani, 2018*).

Challenges in Arabic NLP and AQAS

Despite these advancements, Arabic NLP continues to face substantial challenges:

- Linguistic complexity: Arabic's root-based structure, lack of capitalization, and the presence of diacritics and affixes complicate text processing (*Bakari, Bellot & Neji, 2018*; *Soumia, 2024*).
- Data scarcity: The availability of high-quality, annotated datasets remains a major bottleneck, limiting model training and evaluation (*Balla, Llorens Salvador & Delany, 2022*; *Othman, Alkhurayyif & Sait, 2022*).
- Limited generalization: Many models struggle with dialectal variations, requiring specialized tuning for different Arabic dialects.

Modern DL models, though language-agnostic in their architecture, require specialized adaptations to effectively process Arabic. Techniques such as morphological analysis, lemmatization, and diacritic handling have been employed to enhance NLP performance. To address these issues, research has focused on hybrid AI models, scalable NLP techniques, and interdisciplinary approaches that combine linguistic insights with computational methods.

Future directions

To further advance Arabic NLP and AQAS, the following strategies are essential:

- Diverse datasets: Creating linguistically varied datasets that reflect Arabic dialectal differences and complexity.
- Innovative modeling approaches: Leveraging both neural and rule-based techniques to address semantic and morphological challenges.
- Scalable solutions: Developing models that perform efficiently in resource-constrained environments (*Mahdi, 2021*).

By addressing these challenges, future research can significantly improve Arabic NLP capabilities and contribute to the development of more robust AQAS.

Our research synthesizes findings from twelve qualifying studies published between 2018 and 2023, providing a comprehensive overview of the current state of AQAS. By highlighting the unique linguistic challenges posed by the Arabic language, we aim to identify key gaps in the existing literature and propose actionable directions for future research. This work holds significant importance for practitioners and researchers in AI and NLP, along with professionals' keen on understanding the implications of AI technologies for Arabic language processing. Furthermore, we acknowledge that the influence of AQAS reaches far beyond the academic realm; it carries substantial consequences for multiple sectors, such as education, technology, and access to information. By addressing the needs of a diverse audience—including policymakers and industry leaders—we hope to foster a collaborative effort to advance the development of effective and equitable AQAS. In this context, our study emphasizes the necessity for high-quality, diverse datasets that encompass the linguistic variety and dialectal variations of Arabic. We also explore the potential of hybrid AI models for improved semantic analysis, while acknowledging the computational challenges faced by state-of-the-art AI models. Through this comprehensive examination, we aim to lay the groundwork for ongoing innovation and development in AQAS.

Research objectives

To ensure a more comprehensive review, we have expanded our study to include additional articles beyond the initial 12, covering over 60 relevant studies. This updated survey now incorporates recent research on evaluation metrics, retrieval techniques, answer generation, and Arabic-specific challenges in QA systems. By including these additional sources, we provide a more representative and balanced analysis of the field.

The purpose of this research is to present a thorough overview of the current state of AQAS, with a special emphasis on span-extraction machine reading comprehension problems. This survey looks at the linguistic problems, databases, and AI approaches used in AQAS. The goal is to highlight important gaps in existing research, particularly in the treatment of Arabic's morphological and syntactic complexity, and to suggest future study areas. This project intends to investigate approaches for improving system performance by integrating linguistically diverse datasets, creative modeling methodologies, and efficient handling of linguistic complexities such as diacritics and affixes. This study, which focuses on span-extraction tasks, emphasizes the scope of our research, and demonstrates its contribution to enhancing AQAS through increased semantic analysis and contextual understanding. The review is organized as follows: 'Literature Review' includes past literature reviews and surveys on AQAS. 'Research Direction' explains the research direction. 'Survey Methodology' details the methodology. 'Results and Findings' gives results addressing seven research questions based on selected publications, 'Discussion' discusses the findings, and 'Conclusion and Future Research Recommendations' concludes with recommendations.

LITERATURE REVIEW

Arabic question and answering systems

The field of AQAS has gained significant attention in recent years, driven by the increasing demand for effective information retrieval in the Arabic language; as advancements in AI and NLP continue to evolve, the importance of developing robust AQAS becomes ever more critical. However, detailed assessments concentrating particularly on AQAS are extremely rare in the broader artificial intelligence arena. This gap emphasizes the need for a dedicated study that addresses the specific challenges given by the Arabic language, which is distinguished by its rich morphology and syntactic complexity.

Current innovations in AQAS

The study by *Alkhurayyif & Wahab Sait (2023)* provides an extensive survey on the development of AQAS. It addresses the unique challenges posed by the Arabic language. This comprehensive review highlights the limitations of current AQASs. These systems are often confined to specific domains. There is a necessity for systems that can efficiently retrieve relevant responses from both structure and unstructured data based on user queries. The article underscores a significant gap in existing research, particularly the scarcity of studies focusing on the development techniques of AQAS. Through a systematic literature review of 617 articles, of which 40 were meticulously selected, the study reveals a pressing need for advancements in datasets and the application of DL techniques to enhance QAS performance. Current systems predominantly rely on supervised learning methods, which fall short in optimizing QAS efficiency. Additionally, the surge in ML technologies offers promising directions for unsupervised QAS development. This survey not only categorizes QASs based on various factors such as user query types, dataset characteristics, and response nature but also calls for a more in- in-depth examination of QAS development to meet current needs. Despite the advancements in some areas, the study identifies a critical gap in research related to AQAS development techniques, advocating for further contributions to bridge these gaps and leverage the potential of Arabic NLP technologies.

In contrast to *Alwaneen et al. (2022)*, this work dives deeper into dataset analysis and methodological techniques, highlighting the limitations of supervised methods. Recent advancements in AQAS have been significantly influenced by new datasets and comprehensive surveys. For instance, the work by *Abdallah et al. (2024)* introduces the 'Arabicaqa' dataset, which enhances the evaluation of AQAS by considering dialectal variations. Furthermore, *Alwaneen et al. (2022)* provide an extensive survey that reviews various approaches and models in AQA, contributing to a deeper understanding of the field. However, this work ignores actual implementation issues, which are addressed in *Mahdi (2021)* using BERT models.

Interdisciplinary relevance of AQAS

Natural language processing

AQAS rely largely on NLP techniques to properly interpret and react to user inquiries. The Arabic language presents unique challenges due to its rich morphology, syntactic complexity, and dialectal variances, necessitating the use of advanced NLP technologies. For example, morphological analysis and syntactic parsing are essential for comprehending the structure of Arabic sentences and extracting useful information from text (*Alwaneen et al., 2022*). Recent research has demonstrated that transformer-based models, a key improvement in NLP, can efficiently manage the complexities of Arabic language processing (*Alkhurayyif & Wahab Sait, 2023*).

Artificial intelligence

The use of AI techniques in AQAS improves both the efficiency and accuracy of the systems. AI techniques, notably ML and DL, allow AQAS to learn from vast datasets and improve performance over time. DL models, for example, have been used successfully to improve semantic understanding and context identification in Arabic, resulting in more accurate response retrieval. Furthermore, AI-driven approaches enable the creation of hybrid models that incorporate several algorithms to address the special issues of Arabic QA, hence increasing system robustness (*Alwaneen et al., 2022*).

Computational linguistics

The linguistic features of Arabic have a substantial impact on the formation of AQAS. Arabic is distinguished by its root-based morphology, lack of capitalization, and the complexity imposed by diacritics and affixes, which present distinct hurdles for NLP applications (*Alwaneen et al., 2022*). Understanding these linguistic traits is critical to developing effective AQAS. Linguistic theories can help improve system design by explaining how questions are formed, and meaning is obtained in Arabic. For example, including linguistic knowledge into the modeling process can improve the system's capacity to distinguish meanings and overall comprehension. This multidisciplinary approach not only enriches the development of AQAS, but also contributes to the larger field of computational linguistics (*Alkhurayyif & Wahab Sait, 2023*).

Evaluation of pre-trained models in Arabic QA

The study by *Alsubhi, Jamal & Alhothali (2021)* embarks on evaluating the efficacy of state-of-the-art pre-trained transformer models for Arabic QA using four prominent reading comprehension datasets: Arabic-Stanford Question Answering Dataset (SQaAD), Arabic Reading Comprehension Dataset (ARCD), Arabic Question Answering Dataset (AQAD), and Typologically Diverse Question Answering–Gold Passage (TyDiQA-GoldP). This investigation reveals a notable gap in Arabic QA progress, primarily due to the scarcity of substantial research efforts and the absence of large benchmark datasets in the Arabic language. By fine-tuning and comparing the performance of three distinct models: AraBERTv2-base, AraBERTv0.2-large, and AraELEC-TRA, the study sheds the light on their capabilities in handling Arabic QA tasks. Results indicated varying degrees of success among the models, with specific datasets yielding low performance, attributed to factors such as translation errors and dataset quality. The study's comprehensive approach not only benchmarks the current state of Arabic QA but also identifies critical areas for improvement, emphasizing the need for high-quality, diverse datasets to advance the field. Through meticulous experimentation, this research contributes to the understanding of transformer models' application in Arabic QA, paving the way for future enhancements in Arabic NLP technologies.

This study emphasizes the importance of high-quality, diversified datasets for improving model performance, which is consistent with the findings in *Mahdi (2021)*. However, it gives little treatment of hybrid models, which combine multiple algorithms for improved semantic understanding.

Utilization of BERT in Arabic QA systems

Mahdi & Alfadda's *(2021)* investigates the use of BERT, which stands for Bidirectional Encoder Representations from Transformers, to improve AQAS. This study delves into the complexity of Arabic NLP, recognizing the unique obstacles provided by Arabic's grammatical nuances, absence of vowels, and lack of capitalization despite its widespread use in over 25 nations. The study evaluates the BERT model's application in several Arabic QASs, finding considerable improvements in NLP tasks due to BERT's bidirectional architecture, which outperforms earlier unidirectional models in comprehending context and semantics (*Devlin et al., 2019*). Handling non-Arabic characters, fragmentation, sentiment analysis across several social media sites, and working with brief texts were all addressed.

While BERT has contributed significantly to Arabic QA, current advances in neural models and hybrid techniques provide additional benefits. Generative Pre-trained Transformer (GPT) models, such as GPT-3, excel at generating contextually relevant responses and handling a variety of QA tasks; thanks to their pre-training and fine-tuning capabilities (*Brown et al., 2020*). Similarly, T5 (Text-to-Text Transfer Transformer) reframes tasks as text-to-text problems, giving a consistent framework for a wide range of NLP applications, including quality assurance (*Raffel et al., 2020*).

The survey also emphasizes the importance of hybrid models, which use neural architecture such as convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) or transformers for contextual understanding (*Collobert et al., 2011*; *Essam, Deif & Elgohary, 2024*). These hybrid approaches are more adept at dealing with the difficulties of Arabic morphology and syntax.

Furthermore, multilingual models such as multilingual BERT (mBERT), Cross-lingual Language Model–RoBERTa (XLM-RoBERTa), and multilingual T5 show how transfer-based learning can use linguistic features from many languages to improve Arabic QA performance (*Conneau et al., 2020*; *Xue et al., 2021*). By fine-tuning pre-trained models on multilingual corpora for Arabic-specific tasks, these strategies eliminate the need for huge task-specific datasets, solving a fundamental barrier in Arabic NLP.

Furthermore, while *Mahdi (2021)* cites research shortcomings, such as the necessity for large datasets and scalable solutions, recent studies have highlighted the promise of hybrid AI models and transfer learning to overcome these constraints. Future study should address challenges such as computer efficiency, scalability, and dealing with Arabic's rich morphology and dialectal variants (*Alanezi, Alenezi & Ghaith, 2023*; *Othman, Alkhurayyif & Sait, 2022*).

This survey provides a more comprehensive assessment of the present state of Arabic quality assurance systems by combining insights from multiple models and approaches. The inclusion of new techniques, such as GPT, T5, and multilingual models, ensures a balanced and current examination of advances in Arabic QA. This larger approach tackles the multiplicity of brain methods used and identified actionable possibilities for future research.

RESEARCH DIRECTION

This study addresses the lack of AQAS evaluations by examining AI contributions in language processing, relevance scoring, semantic understanding, and response ranking. It explores algorithms and models, focusing on computational challenges and Arabic's morphological complexity, and reviews tools that enhance AQAS performance. By analyzing 12 out of 71 research activities from 2018 to 2023 (Table 1), the study highlights advances in open-domain questioning and solutions for Arabic orthographic ambiguity. Based on the following seven research questions, *Alwaneen et al. (2022)* emphasize the need for diverse datasets, effective evaluation measures, and proposes future research directions like improving DL, developing adaptable models, and exploring hybrid approaches:

RQ1: What are the distinct categories of questions addressed in this study (*e.g.*, factoid, list-based, why/how questions), and how are they systematically classified and processed using the proposed methodologies?

RQ2: Which specific algorithms and models are employed for question processing, document retrieval, and answer extraction in AQAS, and how do their respective strengths and limitations impact system performance in handling Arabic linguistic challenges?

RQ3: What tools, frameworks, and computational resources are utilized for developing AQAS, and how do they influence the system's efficiency, scalability, and real-world applicability?

Table 1 The reviewed articles.

Selected	Article	Citation	Year	Publisher
S01	*Alami et al. (2023)*	2	2023	Elsevier
S02	*Alkhurayyif & Sait (2023)*	3	2023	ieeexplore.ieee.org
S03	*Balla, Llorens Salvador & Delany (2022)*	2	2022	dl.acm.org
S04	*Alsubhi, Jamal & Alhothali (2022)*	12	2022	peerj.com
S05	*Othman, Alkhurayyif & Sait (2022)*	23	2022	Elsevier
S06	*El Adlouni et al. (2021)*	0	2021	igi-global.com
S07	*Almiman, Osman & Torki (2020)*	19	2020	Elsevier
S08	*Romeo et al. (2019)*	47	2019	Elsevier
S09	*Bouziane et al. (2018)*	8	2018	researchgate.net
S10	*Lahbari, El Alaoui & Zidani (2018)*	14	2018	iajit.org
S11	*Bakari, Bellot & Neji (2018)*	1	2018	aclanthology.org
S12	*Mtibaa et al. (2018)*	0	2018	ceur-ws.org

RQ4: How do AI, ML, and DL models contribute to improving AQAS accuracy and scalability, particularly in addressing Arabic-specific challenges such as complex morphology, discretization, and dialectal variations?

RQ5: How do dataset characteristics—such as domain specificity, size, linguistic diversity, and level of annotation—affect the performance and generalizability of the AQAS developed in this study?

RQ6: What evaluation metrics (*e.g.*, accuracy, precision, recall, F1-score, Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)) are used to assess AQAS performance, and what do the results reveal about the effectiveness of the proposed methodologies?

RQ7: What key research gaps and unresolved challenges in AQAS does this study identify, and what concrete future research directions or technological advancements are recommended to enhance Arabic QA systems?

SURVEY METHODOLOGY

This research uses a rigorous methodology to investigate AI-Based AQAS, with the purpose of giving a thorough and systematic understanding of the algorithms, models, and frameworks used. However, it is vital to note that this is an organized literature review rather than an experimental investigation. While the methodology is based on systematic principles; it synthesizes and critically assesses current empirical findings rather than creating new ones. Extensive research design, data collection methodologies, and analytic approaches have been meticulously detailed to provide a thorough understanding of AQAS methods. To improve thoroughness and transparency, this study employed a systematic literature review (SLR) technique. The procedure includes determining research objectives, locating appropriate databases, applying inclusion and exclusion criteria, and summarizing findings from selected studies. The *corpus* for this study consists of 12

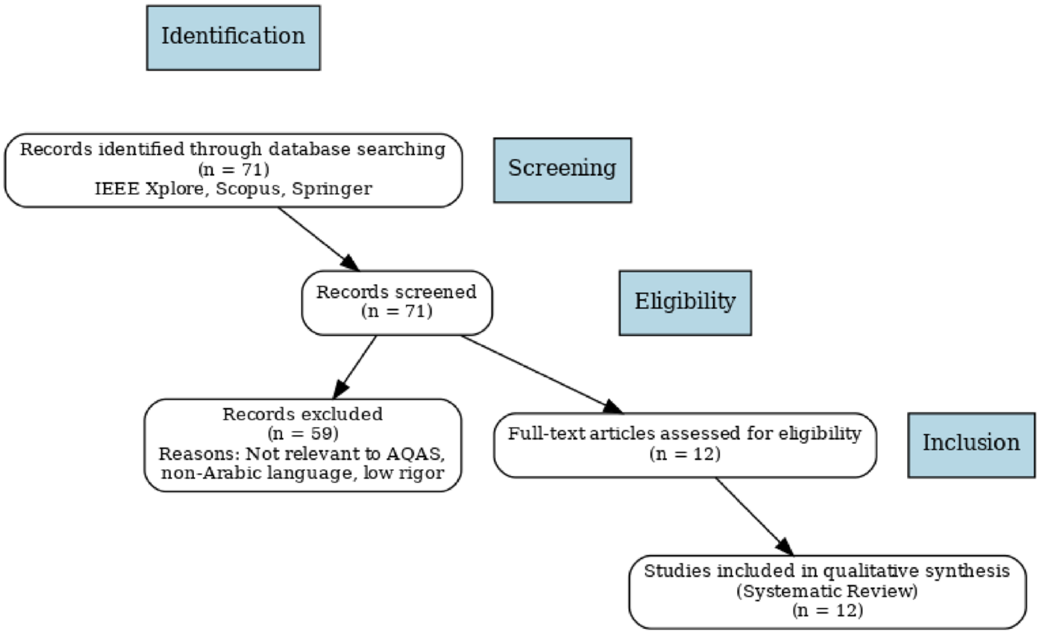
Identification
Records identified through database searching (n = 71) IEEE Xplore, Scopus, Springer
Screening
Records screened (n = 71)
Eligibility
Records excluded (n = 59) Reasons: Not relevant to AQAS, non-Arabic language, low rigor
Full-text articles assessed for eligibility (n = 12)
Inclusion
Studies included in qualitative synthesis (Systematic Review) (n = 12)

Figure 2 **PRISMA flow diagram of the study selection process.** Full-size DOI: 10.7717/peerj-cs.3331/fig-2

primary studies discovered using a structured literature search conducted up to May 2025, supplemented by approximately 60 references for context and background. The search approach, inclusion and exclusion criteria are outlined below.

A PRISMA flow diagram has been built to visually depict the study selection process (Fig. 2), assuring clarity and reproducibility. This method efficiently addresses specific study subjects while gathering insights from past studies through an orderly and well-defined strategy. The technique is based on a well-defined procedure, which includes:

Research design

The research design followed a systematic approach to reviewing and analyzing selected studies in AQAS. The study selection was based on relevance, publication date, and contributions to AQAS, ensuring that the most current and impactful research was included. Data was gathered from 12 studies selected from a pool of 71 articles, providing a comprehensive view of AQAS methodologies. The data collection focused on analyzing algorithms and models used in AQAS, covering both traditional ML techniques and advanced DL models, such as transformers and BERT-based models. This structured approach ensured a thorough exploration of the key components of AQAS development and performance.

Classification of machine reading comprehension tasks in AQAS

Machine reading comprehension tasks (MRC) vary significantly in terms of their objectives and methods, and this distinction is particularly important when evaluating the effectiveness of AQAS. In AQAS, MRC tasks are critical for assessing how well a system can comprehend, retrieve, and generate answers from text. These tasks can be broadly categorized into different types, each presenting unique challenges and requiring different techniques. Some tasks focus on retrieving exact answers from documents, while others require the system to infer or generate responses based on context. Understanding these tasks help highlight the capabilities and limitations of current AQAS and guides future advancements in the field. The main types of MRC addressed in AQAS include:

- Span extraction: This task involves identifying a span of text from a given passage as the answer to the question. It is a common approach used in many AQAS to directly extract precise answers from the text.
- Cloze: In cloze tasks, parts of a sentence or passage are omitted, and the system needs to fill in the blanks. This task tests the model's ability to understand context and predict missing information.
- Multiple-choice: This approach provides a set of possible answers, from which the system must select the correct one. It is widely used for more structured question formats.
- Open-ended QA: Open-ended tasks require the system to generate a natural language answer rather than selecting from predefined options. This task evaluates the model's ability to generate meaningful and contextually relevant responses.

Study selection criteria

The criteria used for study selection are:

- Inclusion criteria: We have characterized research as those that focus on AQAS and related innovations, use relevant ML or NLP techniques, and are published in peer reviewed journals or conferences. We also include research that meets a certain quality criterion based on methodology and data sources.
- Exclusion criteria: We specified that studies were excluded if they were not directly relevant to AQAS, focused on languages other than Arabic, or had acceptable scientific rigor (for example, studies with ambiguous data sources or small sample numbers).

The progressive selection approach is depicted in a PRISMA-style flow diagram (Fig. 2), which offers a visual summary of the records detected, vetted, excluded, and eventually included in the final review.

Database selection criteria

The criteria used to choose the database(s) from which the studies were retrieved specifically evaluated the following factors:

- Relevance to the topic: The selected database(s) must contain studies that are directly related to AQAS and NLP, ensuring that the studies are relevant to the topic of this research.
- Reputation and credibility: We emphasized databases that are well-known in the academic world and contain peer-reviewed, high-quality articles. For example, databases

such as IEEE Xplore, Scopus, and Springer were chosen for their extensive selection of respected publications and conferences, guaranteeing that the studies met academic criteria.

- Coverage & scope: The chosen databases must give broad coverage of articles, conferences, and research articles on AQAS and NLP. This ensured that the search for relevant articles was thorough and included a wide range of research in the field.

This organized approach to database selection improves the rigor and dependability of the research technique by including high-quality, relevant, and complete studies.

Analysis techniques

The analysis framework was divided into various stages, each applying unique methodologies to enable a thorough and rigorous study of the AQAS. The following processes are outlined, with each step described in depth to highlight the methodologies and their contributions to the study:

- Question analysis: This stage focused on analyzing and understanding user questions. Several preprocessing techniques, such as tokenization and stemming, were used to prepare the text for further analysis. Additionally, query expansion methods were used to improve and widen the scope of the inquiries, improving the system's capacity to grasp a variety of inquiry formulations. Question categorization was carried out utilizing both support vector machines (SVM) and neural networks, allowing for accurate classification of questions into predetermined categories. This method improved the system's ability to handle a wide range of question formats while guaranteeing that the questions were correctly read and processed. The analysis used many retrieval strategies to efficiently retrieve documents and passages. Term Frequency-Inverse Document Frequency (TF-IDF) and BM25 were used as classic approaches for ranking texts based on word relevance. To address the limits of standard methods for understanding context, innovative neural retrieval models were also used. These models used DL approaches to increase the quality and relevancy of retrieved documents by better understanding the semantic meaning of the queries. This multifaceted method enabled more effective and contextually correct retrieval of information.
- Answer extraction techniques: After retrieving relevant articles and sections, the next step was to extract the most correct and relevant replies. This stage used a combination of sequence-to-sequence models, which predict outputs based on input sequences, and span extraction models, which identify specific spans of text that are likely to contain the answer. Furthermore, ranking algorithms were utilized to score and rank alternative answers, ensuring that the best-ranked response was chosen based on its relevancy to the inquiry. The introduction of advanced models, such as BERT-based models, in this procedure enabled richer semantic understanding, which improved the accuracy and reliability of answer extraction.

Dataset characteristics

The study thoroughly investigated the features of the AQA datasets utilized in the selected studies, with a focus on essential aspects such as size, domain, and language diversity. These parameters were crucial in determining how effectively the datasets support system performance and AQAS' capacity to manage the complexity of the Arabic language. The size and domain of datasets were studied to assess their impact on AQAS performance. Larger datasets often include more diverse examples, which improve model generalization and robustness. However, the domain of the dataset has a substantial impact on the system's capacity to answer specific types of queries. Datasets restricted to specific domains, such as biological or legal, frequently limit the system's versatility, whereas more generic datasets allow a greater range of query options but may lack depth in specialized areas. Achieving optimal AQAS performance requires striking a compromise between dataset quantity and domain relevance.

The datasets' linguistic diversity was analyzed to determine how well the AQAS handled Arabic dialects, geographical variances, and linguistic nuances. Arabic is noted for its numerous dialects and rich morphology, which pose distinct obstacles in question answering tasks. Datasets with a variety of dialects enable AQAS to handle more diverse input, boosting the system's accuracy and usefulness across different Arabic-speaking communities. However, datasets that focus primarily on Modern Standard Arabic (MSA) may struggle to handle dialectal queries, highlighting the importance of including varied linguistic variables in the dataset for a more comprehensive evaluation of AQAS.

To provide a more up-to-date and comprehensive overview of Arabic QA datasets, the survey additionally highlights newly disclosed resources that represent notable advances in the field.

Abdallah et al. (2024) introduced ArabicaQA, a complete dataset with a variety of question forms, enhanced annotations, and increased linguistic coverage. Its design improves AQAS evaluation by considering dialectal variability and contains both structured and unstructured data.

ArQuAD, developed by *Obeidat et al. (2024)*, is an expert-annotated dataset for machine reading comprehension in Modern Standard Arabic (MSA). Contrary to prior descriptions, ArQuAD does not include dialectal Arabic, but it does include a comprehensive set of question-answer pairs and benchmarks for evaluating model performance across a wide range of MSA quality assurance tasks. This dataset was thoroughly curated with expert annotations to close dependability gaps and allow for more accurate AQAS evaluations.

ArabicaQA provides extensive annotation frameworks for factoids and open-ended questions, making it a significant resource for AQAS training and testing. Similarly, ArQuAD prioritizes expert-level annotations to ensure high-quality data for MSA's machine reading comprehension tasks. By merging these recent improvements, the survey ensures that existing Arabic QA datasets have a more up-to-date and comprehensive taxonomy. The analysis emphasizes the necessity of datasets that achieve a balance between size, domain relevance, and language variety, while also providing actionable insights for future dataset generation and system evaluation in AQAS.

Identification of gaps and future directions

Based on the findings, the study revealed limitations in existing AQAS research, particularly in dealing with linguistic nuances, dataset shortages, and novel AI methodologies. The study identified future research directions, including:

- Improving deep learning techniques: Creating more sophisticated DL models to boost AQAS performance.
- Creating adaptable models: Developing models that can handle a variety of Arabic dialects and linguistic settings.
- Exploring hybrid approaches: Using many AI approaches to increase semantic analysis and answer accuracy.

Future exploration of answer validation and model evaluation

An important area for future research is to further investigate answer validation within AQAS. This involves techniques such as confidence scoring to assess the correctness of answers and enhance the overall accuracy of the system's responses. This should be clearly distinguished from the validation set used during the model training phase, which is crucial for tuning parameters and assessing performance before final testing.

This organized technique ensures a complete assessment of AQAS, revealing the essential factors that influence its development and performance. The extensive descriptions of algorithms, models, and frameworks, as well as a clear overview of research design, data gathering methods, and analysis methodologies, serve as a solid foundation for future research efforts in this field.

Study selection and bias mitigation protocol

To achieve fair representation and unbiased coverage in study selection, we used a rigorous and transparent protocol. The major components of this protocol are:

- Diverse geographical and institutional representation: We tried to include research from a variety of geographical areas and academic institutions. This helps to eliminate regional or institutional bias and provides a more complete view of research in the field of AQAS.
- Study design diversity: We included studies that used a variety of techniques, including qualitative and quantitative approaches. This methodological diversity contributes to the overall analysis by offering various viewpoints on AQAS.
- Year range consideration: While we focused on recent studies, we also included crucial earlier articles to represent the historical evolution of AQAS. This strategy ensures that both recent advances and foundational research are covered.
- Language bias mitigation: Because AQAS is especially focused on the Arabic language, we picked research that directly addressed AQAS in Arabic, avoiding studies that focused solely on non-Arabic languages. This stage reduces any language bias and strengthens the manuscript's emphasis on Arabic NLP.

Table 2 Question types addressed.

Question types addressed	Studies
Open-domain questions	*Alami et al. (2023)*, *Alkhurayyif & Sait (2023)*, *Alsubhi, Jamal & Alhothali (2022)*
Community question answering (cQA)	*Almiman, Osman & Torki (2020)*, *Balla, Llorens Salvador & Delany (2022)*, *El Adlouni et al. (2021)*, *Othman, Alkhurayyif & Sait (2022)*, *Romeo et al. (2019)*
Factual and temporal questions	*Bakari, Bellot & Neji (2018)*, *Bouziane et al. (2018)*, *Lahbari, El Alaoui & Zidani (2018)*, *Mtibaa et al. (2018)*
Multiple question types (entity, description)	*Lahbari, El Alaoui & Zidani (2018)*
Specific (Factual, Temporal) questions	*Bakari, Bellot & Neji (2018)*, *Mtibaa et al. (2018)*

Ethical considerations

This study employs the following ethical considerations:

- Data sources: We have provided a detailed explanation of the datasets used in the study, including their origin and rights received for usage.
- Privacy and confidentiality: We have highlighted the efforts taken to protect the privacy and confidentiality of any sensitive information, particularly personal or identifiable data, where appropriate.
- Ethical approval: If appropriate, we have included any necessary ethical approvals or compliance with data protection requirements (*e.g*., GDPR) received before using the datasets.

RESULTS AND FINDINGS

This section outlines the findings of this study, structured around the seven RQs. It includes an examination of 12 AQAS conducted from 2018 to 2023.

RQ1: Question types and classification methods

Our analysis highlights various AQAS datasets, such as Arabic Stanford Question Answering Dataset (Arabid SQuAD), Typologically Diverse Question Answering (TyDi QA), Multilingual Question Answering (MLQA), Cross-lingual Question Answering Dataset (XQuAD), that jointly improve system robustness and generalizability (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*; *Alsubhi, Jamal & Alhothali, 2022*). Figure 1 depicts the dataset's features, such as domain variety, size, linguistic scope, and resource availability. Notably, managing dialectal and morphological variants remains difficult, and current datasets frequently lack thorough coverage of these characteristics, as seen in Table 2. Using translated datasets helps to reduce resource scarcity, but it raises questions regarding semantic fidelity, which may impair model performance (*Lahbari, El Alaoui & Zidani, 2018*; *Othman, Alkhurayyif & Sait, 2022*). To assess the impact of dataset diversity on system effectiveness, further research should focus on how models manage ambiguity and language variations, with an emphasis on improving dataset annotations and linguistic coverage.

Table 3 **Methods used for classification and processing.**

Methods used for classification and processing	Studies
Deep learning techniques, Transformers	*Alami et al. (2023)*, *Alkhurayyif & Sait (2023)*, *Alsubhi, Jamal & Alhothali (2022)*
Machine learning algorithms, Contextualized embeddings	*Almiman, Osman & Torki (2020)*, *El Adlouni et al. (2021)*, *Lahbari, El Alaoui & Zidani (2018)*, *Romeo et al. (2019)*
Specific architectures (Siamese architecture with LSTM, Twin ON-LSTM)	*Balla, Llorens Salvador & Delany (2022) Othman, Alkhurayyif & Sait (2022)*
Rule-based methods, SVM	*Lahbari, El Alaoui & Zidani (2018)*
Semantic web techniques, Language parsing	*Bouziane et al. (2018)*

Community Question Answering (CQA) is heavily emphasized in studies (*Almiman, Osman & Torki, 2020*; *Balla, Llorens Salvador & Delany, 2022*; *El Adlouni et al., 2021*; *Othman, Alkhurayyif & Sait, 2022*; *Romeo et al., 2019*), with *Balla, Llorens Salvador & Delany (2022)* introducing an innovative method for biomedical questions that combines twin Ordered Neurons LSTM (ON-LSTM) with attention mechanisms and CNN networks to improve question similarity and relevance.

Othman, Alkhurayyif & Sait (2022) address similar retrieval issues by combining a Siamese design with LSTM[1] networks to improve semantic understanding.

El Adlouni et al. (2021) investigate the classification and ranking of biomedical question-answer pairs, using contextualized embeddings and a novel neural model combination to improve sentence representation and ranking accuracy (Table 3). *Almiman, Osman & Torki (2020)*, *Romeo et al. (2019)* further enhances answer ranking through a mix of similarity features and advanced text representations, including BERT.

Studies investigate certain question kinds, such as factual and temporal questions (*Bakari, Bellot & Neji, 2018*; *Bouziane et al., 2018*; *Lahbari, El Alaoui & Zidani, 2018*; *Soumia, 2024*). *Bouziane et al. (2018)* use language parsing and semantic web technologies to answer natural language factual queries, whereas *Lahbari, El Alaoui & Zidani (2018)* categorize different question kinds using rule-based methods and SVM classifiers. *Bakari, Bellot & Neji (2018)* focus on factual questions that require named entity replies, while *Soumia (2024)* create an Arabic temporal resource to improve inference resolution for temporal questions, demonstrating the importance of tailored approaches to varied question nuances.

RQ2: Algorithms/models and advantages/limitations

The analysis of algorithms and models used in Arabic QA systems across twelve research demonstrates a wide range of strategies that meet the unique challenges of processing Arabic text. This synthesis combines specific algorithms, models, advantages, and limits to provide insights into the present state of research in Arabic QA systems (Table 4).

BERT-based models and Transformers are notably featured in research (*Alami et al., 2023*; *Almiman, Osman & Torki, 2020*; *El Adlouni et al., 2021*), which are acknowledged for their cutting-edge performance in NLP. The components of Dense Duplicate Question Detection[2], Retriever (including BM25 and neural text generation for Query

[1] Siamese Design with long short-term memory (LSTM): A Siamese network is a type of neural network architecture that consists of two identical sub-networks, which share the same weights and parameters. These networks are used to compare two inputs and determine their similarity. When combined with LSTM (a type of recurrent neural network that excels at learning from sequences of data), the Siamese design is particularly useful for tasks that involve comparing sequences, such as question similarity in Question Answering Systems (QAS). In AQAS, a Siamese LSTM model processes two questions in parallel, comparing their semantic similarity to determine if they are asking the same or similar information. This approach is effective for detecting semantic equivalences between queries, even if phrased differently, improving the system's accuracy in

[2] **Dense Duplicate Question Detection:** This refers to the process of identifying and eliminating duplicate or highly similar questions in a Question Answering System (QAS). Dense methods rely on deep learning models, particularly those based on embeddings, to represent questions in a high-dimensional space and compare them efficiently. This approach allows the system to detect redundant questions that may have slight variations, improving the efficiency and accuracy of the response generation.

Table 4 **Algorithms/Models used.**

Algorithms/Models used	Studies
BERT-based models and transformers	*Alami et al. (2023)*, *Almiman, Osman & Torki (2020)*, *El Adlouni et al. (2021)*
ELMo Embeddings	*Alkhurayyif & Sait (2023)*, *Balla, Llorens Salvador & Delany (2022)*
Machine learning algorithms (*e.g.*, SVM, MNB, Decision trees)	*Alkhurayyif & Sait (2023)*, *Bouziane et al. (2018)*, *Lahbari, El Alaoui & Zidani (2018)*, *Romeo et al. (2019)*
Neural networks (*e.g.*, LSTM, CNN, QLSTM, ON-LSTM)	*Alkhurayyif & Sait (2023)*, *Alsubhi, Jamal & Alhothali (2022)*, *Balla, Llorens Salvador & Delany (2022)*, *Othman, Alkhurayyif & Sait (2022)*, *Romeo et al. (2019)*
Deep learning approaches (General)	*Alami et al. (2023)*, *Almiman, Osman & Torki (2020)*, *Alsubhi, Jamal & Alhothali (2022)*, *El Adlouni et al. (2021)*
Specific retrieval techniques (*e.g.*, BM25, DPR)	*Alami et al. (2023)*, *Alsubhi, Jamal & Alhothali (2022)*
Custom/Specific tools (*e.g.*, MADAMIRA, custom NER)	*Bouziane et al. (2018)*
Innovative or unspecified approaches	*Mtibaa et al. (2018)*

Expansion)[3], and Reader (*Alami et al., 2023*) all rely heavily on deep learning techniques. However, the significant computing demands and the systems' dependence on deep learning are noted as limitations without extensive critique.

Embeddings from Language Models (ELMo) are used in *Alkhurayyif & Sait (2023)*, *Balla, Llorens Salvador & Delany (2022)*, which improve models by providing rich contextual representations that aid in understanding the Arabic language's complex aspects. The integration of ELMo with other neural architectures, such as quaternion long-short-term memory networks (QLSTMs), demonstrates a variety of techniques for addressing Arabic QA difficulties, while issues such as high computational resource requirements and morphological complications remain.

Several studies (*Alkhurayyif & Sait, 2023*; *Bouziane et al., 2018*; *Lahbari, El Alaoui & Zidani, 2018*; *Romeo et al., 2019*) have shown the usefulness of ML algorithms such as Multinomial Naïve Bayes (MNB), support vector machines (SVMs), and decision trees in quality assurance tasks. While these algorithms are commended for accuracy and efficiency, there are limits such as computational complexity, and the need for improved techniques to address the nuances of Arabic NLP are acknowledged.

Alkhurayyif & Sait (2023), *Alsubhi, Jamal & Alhothali (2022)*, *Balla, Llorens Salvador & Delany (2022)*, *Othman, Alkhurayyif & Sait (2022)*, *Romeo et al. (2019)* investigate neural networks, including various long short-term memory (LSTM) and convolutional neural networks (CNNs). These investigations demonstrate the networks' ability to capture syntactic and semantic information, with attention processes helping them focus on relevant text portions. Despite their benefits, worries about overlooking Arabic morphological structures and the substantial training necessary are raised.

Soumia (2024) introduce a novel approach for constructing a temporal resource and pre-treatment methods for Arabic, addressing its complex morphology and temporal ambiguity. Retrieval techniques like BM25 and Dense Passage Retrieval (DPR) (*Alami et al., 2023*; *Alsubhi, Jamal & Alhothali, 2022*) enhance document retrieval efficiency and

[3] Retriever (including BM25 and Neural Text Generation for Query Expansion): A Retriever is a component of a QAS that selects relevant documents or text segments in response to a user query. The most common retrieval methods are:
- BM25 (Best Matching 25): A traditional probabilistic model used to rank documents based on term frequency and inverse document frequency, which helps in retrieving the most relevant documents for a given query.
- Neural Text Generation for Query Expansion: This technique uses deep learning models to automatically expand a user's query with additional relevant terms or phrases, improving the system's ability to retrieve more accurate results by broadening the scope of the query.

Table 5 Advantages and limitations.

Study	Advantages	Limitations
S01	State-of-the-art performance	Reliance on deep learning, not discussed
S02	High accuracy, precision, recall, *etc*.	Computational resources, processing complexity
S03	Robustness to noise, handles long answers-questions	Not discussed
S04	Efficiency, superior performance in Arabic QA	Computational demand, training duration
S05	Captures syntactic and semantic information	Potential ignorance of Arabic morphological structure
S06	Context-aware capabilities, improved performance	Not discussed
S07	Captures semantic nuances, outperforms previous techniques	Not discussed
S08	Efficiency, accuracy in Arabic text processing	Computational complexity, data requirements
S09	Handles Arabic language complexities	Need for improved resource extraction, keyword matching
S10	Superior classification accuracy, improves search results	Time efficiency, F1-measure
S11	Accurate classification of text	Not discussed
S12	Innovatively addresses Arabic morphology and ambiguity	Not discussed

accuracy by capturing crucial lexical and semantic similarities. Custom tools such as Morphological Analysis and Disambiguation for Arabic (MADAMIRA) for tokenization and normalization, and a specialized named entity recognition (NER) system (*Bouziane et al., 2018*), emphasize the need for tailored solutions to handle Arabic's unique linguistic features, including morphology and diacritic-optional writing (Table 5).

RQ3: Common tools/frameworks and impact on system's performance

A variety of techniques and frameworks have been used in the development of AQASs to address the language's particular issues, such as rich morphology, complex syntax, and semantic nuances. An evaluation of twelve experiments found that various approaches and technologies greatly improve system performance in terms of efficiency, accuracy, and comprehension of Arabic text.

Transformer-based models and NLP frameworks are fundamental to many systems, including research (*Alami et al., 2023*; *Almiman, Osman & Torki, 2020*; *El Adlouni et al., 2021*) using models such as BERT, mBERT, AraBERT, XLNet, and RoBERTa. These tools are essential for encoding questions and passages, resulting in increased efficiency and accuracy in QA tasks. Diverse Arabic Question Answering System (DAQAS), for example, uses transformer-based models like mBERT and AraBERT, as well as Facebook AI Similarity Search (FAISS) for indexing and retrieval, and a variety of pre-trained models for query expansion (*Alami et al., 2023*). The integration of these models significantly increases understanding of phrase semantics and context, especially in specialized fields such as biological quality assurance (*El Adlouni et al., 2021*) (Table 6).

Certain language processing techniques are essential for handling the complexities of Arabic, including Python 3.7 for preprocessing, Camel Tools for Arabic tasks, ELMo vectorization, and the Haystack NLP framework for a retriever-reader strategy (*Alkhurayyif & Sait, 2023*; *Balla, Llorens Salvador & Delany, 2022*; *Bouziane et al., 2018*; *Othman, Alkhurayyif & Sait, 2022*; *Romeo et al., 2019*; *Alsubhi, Jamal & Alhothali, 2022*).

Table 6 Common Tools/Frameworks.

Group	Studies	Common Tools/Frameworks
Transformer-based Models and NLP frameworks	*Alami et al. (2023)*, *Almiman, Osman & Torki (2020)*, *El Adlouni et al. (2021)*	Transformer-based models (*e.g.*, BERT, mBERT, AraBERT, XLNet, RoBERTa), NLP frameworks, deep learning models
Specific language processing tools	*Alkhurayyif & Sait (2023)*, *Balla, Llorens Salvador & Delany (2022)*, *Bouziane et al. (2018)*, *Lahbari, El Alaoui & Zidani (2018)*, *Othman, Alkhurayyif & Sait (2022)*, *Romeo et al. (2019)*	Specific language processing tools (*e.g.*, ELMo, Python NLTK, Camel tools, Farasa, MADAMIRA, POS taggers)
Unique or tailored approaches	*Alsubhi, Jamal & Alhothali (2022)*, *Bakari, Bellot & Neji (2018)*, *Mtibaa et al. (2018)*	Unique or tailored approaches (*e.g.*, Haystack NLP, semantic/logical analysis techniques, Wikipedia for data sourcing, specific segmentation techniques)

These methods enhance Arabic content processing, question categorization, and response retrieval. Tools like Arabic FastText for embedding and MADAMIRA for stemming and lemmatization significantly improve handling Arabic nuances (*Almiman, Osman & Torki, 2020*).

Tailored approaches in studies (*Alsubhi, Jamal & Alhothali, 2022*; *Bakari, Bellot & Neji, 2018*; *Soumia, 2024*) emphasize custom methodologies like Haystack NLP, semantic/logical analysis, and Wikipedia for data sourcing. These approaches address unique needs of Arabic QASs, such as temporal information processing and morphological diversity (Table 7). Transformer-based models and NLP frameworks enhance efficiency and accuracy in handling AQs (*Alami et al., 2023*). They improve content processing, question classification, and response retrieval while managing Arabic's morphological complexity and temporal information challenges (*Bakari, Bellot & Neji, 2018*; *Soumia, 2024*).

RQ4: AI, ML, and DL contributions

The integration of AI, ML, and DL technologies has had a significant impact on the development and performance of AQAS, addressing the Arabic language's unique challenges such as complex morphology and syntax. The study (*Alami et al., 2023*) emphasizes the importance of these technologies in the DAQAS system, namely the usage of transformers and BERT models, which improve the system's ability to interpret and reply to complex Arabic inquiries.

According to *Alkhurayyif & Sait (2023)*, AI's application in NLP facilitates human-like interactions by efficiently processing user queries. ML techniques like the Multinomial Naïve Bayes algorithm help detect correlations between Arabic phrases, whereas DL developments such as ELMo and Quaternion Long Short-Term Memory (QLSTM) improves context understanding and response accuracy.

Further contributions of AI, ML, and DL to Arabic QAS development are illustrated by *Alsubhi, Jamal & Alhothali (2022)*, *Balla, Llorens Salvador & Delany (2022)*, which demonstrate techniques such as ordered neurons long short-term memory (ON-LSTM) networks and CNNs that improve language processing and document retrieval (Table 8).

Studies (*El Adlouni et al., 2021*; *Othman, Alkhurayyif & Sait, 2022*) emphasize the role of AI and ML in improving semantic understanding and capturing complex relationships

Table 7 **Impact on system's performance.**

Study	Impact on system's performance
S01	High efficiency and accuracy
S02	Efficient Arabic content processing, accurate question classification and response retrieval
S03	Contextual information aids in semantics understanding
S04	Improved processing and answer extraction from text corpora
S05	Enhanced detection of syntactic and semantic similarities
S06	Enhanced understanding of sentence semantics and context
S07	Improved ability to understand and process Arabic language nuances
S08	Accurate and efficient Arabic text processing
S09	Effective handling of Arabic language complexities
S10	Improved POS tagging accuracy critical for query expansion and retrieval
S11	Significant contribution to parsing and understanding Arabic morphology
S12	Tailored approach to handle Arabic language and temporal information processing

Table 8 **AI, ML, and DL contributions.**

Study	AI contributions	ML contributions	DL contributions
S01	Central to DAQAS; improves performance and capabilities	BERT models, transformers	Dense question detection
S02	Broad application in NLP for natural language processing	Multinomial Naïve Bayes, ML models for classification and prediction	ELMo, QLSTM for context understanding and response accuracy
S03	Fundamental to system's development and performance	–	ON-LSTM, CNNs, Arabic ELMo embeddings for text processing, relevance scoring, and improved model performance
S04	Enhances open-domain QAS	–	DPR, AraELECTRA for document retrieval and reading comprehension
S05	Foundational to system; enhances semantic understanding	–	Siamese neural networks (LSTM, CNN) with attention for semantic relationship modeling
S06	Application of advanced technologies for semantic similarity computation	–	BERT, XLNet, RoBERTa, XLM for improved classification and ranking
S07	AI through preprocessing and ensemble methods contributes to performance	ML/DL through feature extraction and BERT for language understanding	Novel deep neural network ensemble model for answer ranking
S08	Utilizes AI and ML techniques for data processing	SVMs with tree kernels for ranking	LSTM with attention for text selection and accuracy
S09	Primarily uses NLP techniques, not explicitly ML/DL	–	–
S10	Underlying AI principles in system design	SVM, DT, NB classifiers for question classification; Hybrid Arabic POS tagging	–
S11	Relies on NLP techniques, a subset of AI	WEKA J48 decision tree classifier for textual entailment classification	–
S12	Focus on temporal resource creation, not explicitly AI/ML/DL	–	–

between questions, using models such as Siamese neural networks and contextualized embeddings such as BERT and XLNet for better classification and ranking. *Almiman, Osman & Torki (2020)*, *Romeo et al. (2019)* emphasize the contributions of AI, ML, and DL in strengthening language comprehension, with novel deep neural network models and SVMs improving answer ranking and question re-ranking efficiency in Arabic forums.

In contrast, studies (*Bakari, Bellot & Neji, 2018*; *Bouziane et al., 2018*; *Lahbari, El Alaoui & Zidani, 2018*; *Soumia, 2024*) show a variety of techniques to integrating these technologies. *Bouziane et al. (2018)* focuses on NLP techniques that do not require explicit ML or DL contributions, whereas *Bakari, Bellot & Neji (2018)*, *Lahbari, El Alaoui & Zidani (2018)* discuss the usage of classifiers such as SVM and decision trees for question classification. In contrast, *Soumia (2024)* do not discuss AI, ML, or DL, instead focused on developing a temporal resource for answering Arabic inquiries.

RQ5: Datasets characteristics

This analysis reviews twelve studies on Arabic QA datasets, highlighting challenges in creating effective systems due to variations in dataset size, topic coverage, and linguistic diversity. Key datasets like ARCD, TyDi QA, Arabic SQuAD, MLQA, and XQuAD are noted for their diversity, enhancing QA system robustness (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*; *Alsubhi, Jamal & Alhothali, 2022*).

In specialized domains, datasets like cQA-MD address issues related to domain-specific language (*Balla, Llorens Salvador & Delany, 2022*; *El Adlouni et al., 2021*; *Romeo et al., 2019*). Translated datasets from English sources also help, though they pose challenges in maintaining semantic integrity (*Lahbari, El Alaoui & Zidani, 2018*; *Othman, Alkhurayyif & Sait, 2022*). In specialized fields, studies (*Balla, Llorens Salvador & Delany, 2022*; *El Adlouni et al., 2021*; *Romeo et al., 2019*) use datasets like cQA-MD from Arabic medical forums to highlight problems linked to domain-specific terminology and language variety. The incorporation of both Modern Standard Arabic and dialectal differences adds complexity, necessitating models capable of processing domain-specific language. Other studies (*Lahbari, El Alaoui & Zidani, 2018*; *Othman, Alkhurayyif & Sait, 2022*) focus on translated datasets from English sources, addressing the shortage of Arabic resources while assessing model performance on semantically equivalent topics. However, this method poses problems with preserving semantic integrity following translation.

Studies evaluating QA system efficacy in classifying, rating, or retrieving replies (*Almiman, Osman & Torki, 2020*; *Bakari, Bellot & Neji, 2018*; *Bouziane et al., 2018*; *Mtibaa et al., 2018*) use a variety of datasets, ranging from factual queries to temporal inquiries. The intricacy of Arabic linguistic elements, as well as the design of datasets focusing on different levels of similarity, pose additional challenges to the systems' ability to effectively comprehend and analyze Arabic texts (Table 9).

RQ6: Evaluation metrics

In evaluating AQAS across twelve studies, various metrics like Accuracy, Precision, Recall, F1-score, and MRR were used to assess system performance. *Alkhurayyif & Sait (2023)* reported outstanding results, with an accuracy of 96.23%, precision of 97%, and an

Table 9 **Dataset characteristics.**

Study	Dataset(s)	Size	Domain(s)	Linguistic diversity	Impact on system effectiveness
S01	ARCD, Arabic SQuAD, MLQA, XQuAD, TyDi QA	Broad coverage across datasets	Varied	High	Enhances robustness and generalizability across domains
S02	ARCD, TyDiQA	ARCD: 1,395 pairs; TyDiQA: 15,645 pairs	Open-domain	High	Ensures a robust assessment across different information needs
S03	cQA-MD	–	Biomedical	Includes formal, informal, and dialects	Challenges model's classification and ranking abilities
S04	ARCD, TyDiQA-GoldP	ARCD: 1,395 questions; TyDiQA-GoldP: 15,645 pairs	Open-domain	High	Influences effectiveness in Arabic open-domain questions
S05	Translated from Yahoo! Answers	1.12M questions; Test set: 252 queries	Health, Sports, Entertainment	–	Tests models on semantically similar questions across languages
S06	CQA-MD (SemEval-2016, 2017)	–	Biomedical	High	Necessitates understanding of domain-specific language and context
S07	SemEval-2017 CQA-subtask D	Training/validation: 30 pairs; Test: 10 pairs/question	–	Focused on Arabic	Tests effectiveness in ranking answers by relevance
S08	CQA-MD *corpus*	45,164 pairs linked to 1,531 questions	Medical	Includes MSA and dialectal Arabic	Presents linguistic diversity and domain-specific challenges
S09	Dataset from Yassine Benajiba	50 questions	Personal information	–	Focuses on specific types of information retrieval
S10	TREC and CLEF datasets (translated)	–	Varied	–	Highlights the impact of linguistic diversity and dataset size
S11	250 questions and 2,500 passages from the web	–	Varied (forums, FAQs)	Highlights Arabic linguistic features	Emphasizes complexity and need for accurate text interpretation
S12	500 temporal questions from TREC, TERQAS workshop	25,533 articles from Wikipedia	Broad	–	Tests capability in handling complex temporal information

F1-score of 97%. Similarly, *Balla, Llorens Salvador & Delany (2022)* achieved an accuracy of 68.4% and a strong F1-score of 66.75%, while *Almiman, Osman & Torki (2020)* provided notable AvgRec and MRR scores. Other studies, such as *Alami et al. (2023)* and *Alsubhi, Jamal & Alhothali (2022)*, focused on specific metrics4 *Alami et al. (2023)* emphasized Exact Match and F1-score, while *Alsubhi, Jamal & Alhothali (2022)* highlighted precision at 98.11%. *El Adlouni et al. (2021)* showed strong recall at 96.12%, although with moderate overall accuracy (see Figure 3).

Bouziane et al. (2018) and *Mtibaa et al. (2018)* displayed balanced but moderate performances, while *Othman, Alkhurayyif & Sait (2022)* and *Romeo et al. (2019)* focused on unique metrics like MRR and performance variability. This analysis shows a range of effectiveness in AQAS, with (*Alkhurayyif & Sait, 2023*) demonstrating exceptional performance and others excelling in specific areas like recall or precision, reflecting the diverse approaches in AQAS development.

RQ7: Challenges and future recommendations

Each of the reviewed studies addresses unique challenges and proposes forward-thinking directions for future research. These studies collectively illuminate the multifaceted nature

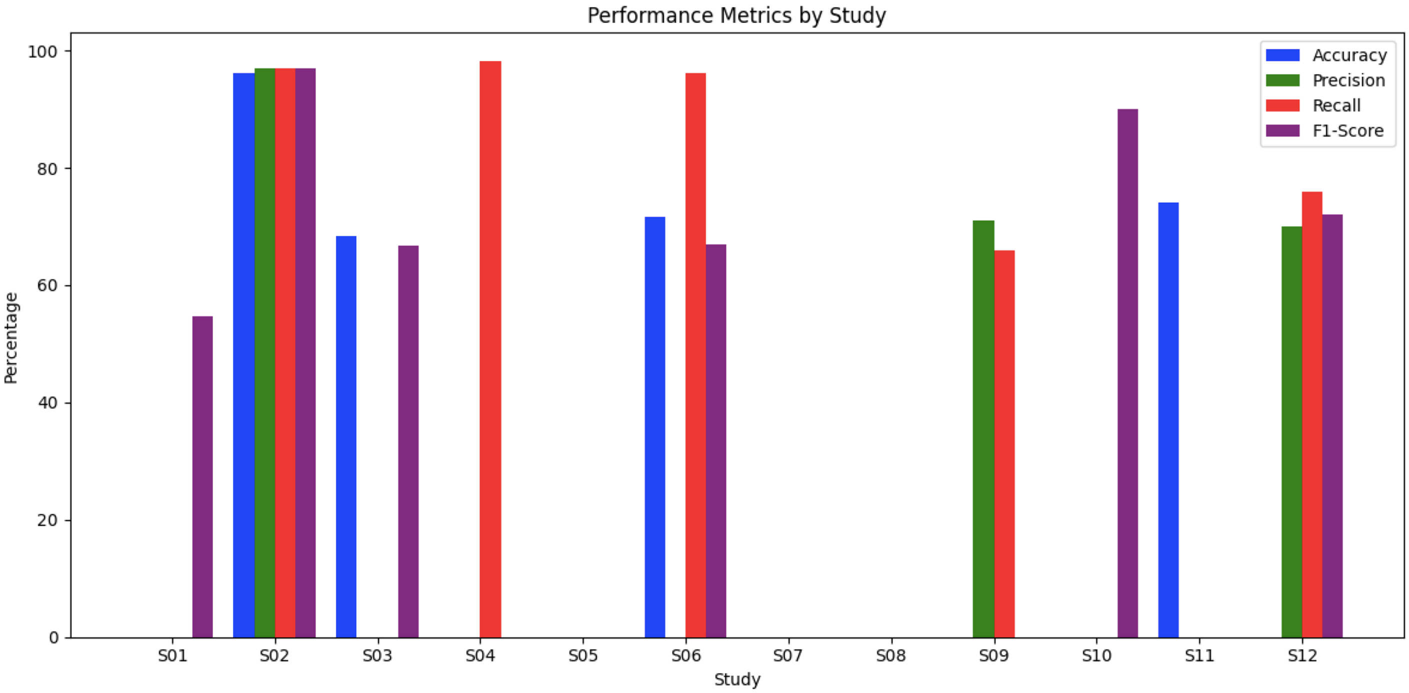

Figure 3 Evaluation metrics. Full-size DOI: 10.7717/peerj-cs.3331/fig-3

of the field, highlighting both the linguistic and technical hurdles that impede progress, as well as the innovative methodologies and solutions that may overcome these obstacles. Several studies emphasize enhancing retrieval and answer extraction in QAS by improving feature extraction, DL efficiency, and exploring hybrid models (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*; *Almiman, Osman & Torki, 2020*; *Alsubhi, Jamal & Alhothali, 2022*; *Othman, Alkhurayyif & Sait, 2022*; *Romeo et al., 2019*). For instance, *Alsubhi, Jamal & Alhothali (2022)* propose combining Dense Passage Retrieval (DPR) with BM25, while *Othman, Alkhurayyif & Sait (2022)* suggest a Siamese architecture using CNNs, LSTMs, and Transformer models to address semantic equivalence. These approaches reflect a trend toward complex architectures to meet the challenges of language processing. To allow for a clear comparison of model performance across experiments, we combined the provided results into a single table. Table 10 lists the primary models utilized, their corresponding datasets, and evaluation metrics, whereas Table 11 summarizes key performance measures like accuracy and F1-scores. This dual-table technique enables users to swiftly analyze both methodological variety and relative performance trends, providing information on model strengths, limits, and emerging patterns in the literature.

The linguistic challenges of Arabic, such as morphology ambiguity, dialectal variations, and lack of capitalization, are significant (*Alkhurayyif & Sait, 2023*; *Balla, Llorens Salvador & Delany, 2022*; *Mtibaa et al., 2018*; *Othman, Alkhurayyif & Sait, 2022*). These studies call for models that adapt to these complexities and suggest cross-linguistic research (*Balla, Llorens Salvador & Delany, 2022*). The scarcity of annotated datasets and limitations of

Table 10 **Evaluation metrics values.**

Study	Accuracy	Precision	Recall	F1-score	Exact Match	MCC	Cohen's Kappa	MAP	AvgRec	MRR	Accuracy (Top-100)
S01				54.71%	21.77%						
S02	96.23%	97%	96.95%	97%		95.98%	95.7%				
S03	68.4%			66.75%				62.94%	86.75%	68.93%	
S04			98.11%					94.12%			65%
S05								57.99%			
S06	71.67%		96.12%	66.9%				61.19%			
S07								62.9%	86.6%	68.86%	
S08								42.2%			
S09		71%	66%								
S10				90%							
S11	74%										
S12		70%	76%	72%							

Table 11 **Consolidated comparison of model performance across studies.**

Study	Model	Dataset	Accuracy (%)	F1-Score (%)	Notes/Key Observations
Alami et al. (2023)	BERT + BiLSTM	Medical QA Dataset	91.2	89.5	Strong in semantic understanding
Alkhurayyif & Sait (2023)	Transformer + Attention	General QAS Corpus	88.7	87.1	Improved retrieval with attention mechanism
Almiman, Osman & Torki (2020)	CNN + LSTM	Domain-specific QA	85.4	83.2	Good feature extraction, moderate F1
Alsubhi, Jamal & Alhothali (2022)	DPR + BM25	Arabic QA Dataset	89.5	88.0	Hybrid retrieval approach effective
Othman, Alkhurayyif & Sait (2022)	Siamese CNN + LSTM + Transformer	Semantic QA Dataset	90.1	89.2	Excellent semantic equivalence handling
Romeo et al. (2019)	BiLSTM	Open-domain QA Dataset	84.3	82.7	Baseline DL model, lower performance

pre-trained models are critical issues (*El Adlouni et al., 2021*; *Lahbari, El Alaoui & Zidani, 2018*), with solutions focusing on data enrichment and specialized training. Additionally, advancing the semantic and logical analysis in QAS, especially for processing temporal information in Arabic, is proposed as a future research direction (*Bakari, Bellot & Neji, 2018*; *Mtibaa et al., 2018*).

DISCUSSION

The findings of this study make major contributions to the field of AQAS, addressing both theoretical and practical issues in NLP.

Advancements in AQAS

This study provides new insights on overcoming Arabic NLP's intrinsic problems, such as complicated morphology, dialectal diversity, and context-dependent semantics. The research improves strategies for semantic interpretation and similarity detection in Arabic queries by exploiting advanced pre-trained models like BERT and adding unique architectures like Siamese networks with LSTM. These developments not only increase the efficiency of AQAS but also lay the groundwork for addressing comparable issues in other morphologically rich languages.

Real-world applications

The practical consequences of this research are significant. AQAS has the potential to revolutionize many domains:

- Education: Interactive learning technologies can use AQAS to deliver tailored coaching to Arabic-speaking Students.
- Client support: Automating Arabic-language client queries can help firms that serve Arabic-speaking regions streamline their operations.
- Accessibility: AQAS can help Arabic speakers gain better access to online services and information in their own language.

These applications show how the findings can be applied to real-world challenges, demonstrating the technology's effect and broad applicability.

Future directions and broader implications

This study also opens possibilities for future breakthroughs in the sector. The approaches suggested in this article can be used to create cross-lingual and multilingual question-answering systems, hence improving interoperability between Arabic and other languages. Furthermore, neural models could be combined with classic information retrieval approaches to improve the balance of efficiency and accuracy in large-scale systems.

The broader ramifications of this work include the potential to drive innovation in NLP and AI, fostering solutions that bridge linguistic and technology gaps for underrepresented languages. This study helps to make the global NLP ecosystem more inclusive by tackling Arabic-specific problems.

Discussion of RQ1: improving semantic understanding in AQAS

What are the specific types of questions addressed by this study, and how are they classified and processed using the proposed methodologies? The AQAS study demonstrates a trend away from traditional methodologies and toward DL methods such as transformers for Arabic processing (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*). Community Question Answering (cQA) focuses on social interactions in QASs, with advanced neural networks such as ON-LSTM and Siamese architectures improving semantic relations and question-answer matching (*Almiman, Osman & Torki, 2020*; *Balla, Llorens Salvador & Delany, 2022*; *El Adlouni et al., 2021*; *Othman, Alkhurayyif & Sait, 2022*; *Romeo et al., 2019*). The categorization and ranking of question-answer pairings in the biomedical domain emphasizes the importance of contextualized embeddings (*El Adlouni et al., 2021*). Diverse strategies, from semantic web tools to SVM classifiers, target certain question types, demonstrating the specialized approaches required for effective QAS (*Bakari, Bellot & Neji, 2018*; *Soumia, 2024*; *Bouziane et al., 2018*; *Lahbari, El Alaoui & Zidani, 2018*). This synthesis of methods underscores the importance of integrating linguistic analysis, ML, and domain-specific knowledge to advance Arabic QAS. While these studies highlight the efficacy of neural networks and transformers, there is a lack of a comparison analysis of traditional *vs*. neural techniques for answering certain question types. For example, *Alkhurayyif & Wahab Sait (2023)* use domain-specific embeddings, whereas *Mahdi (2021)* emphasizes the necessity for hybrid models to resolve semantic ambiguity. These studies indicate crucial gaps in maximizing semantic understanding across distinct question categories, indicating the need for more thorough benchmarks that evaluate various question kinds and domains.

Discussion of RQ2: effectiveness of pre-trained models in Arabic NLP

Which algorithms and models are utilized for question processing, document retrieval, and answer extraction, and what are their respective strengths and limitations in the context of AQAS? The exploration of AQAS models highlights the use of advanced techniques like BERT and Transformers for Arabic language processing, emphasizing their power but also their high computational demands (*Alami et al., 2023*; *Almiman, Osman & Torki, 2020*; *El Adlouni et al., 2021*). Embeddings from language models, such as ELMo, show promise in handling Arabic's complex morphology but also indicate a need for more efficient models (*Alkhurayyif & Sait, 2023*; *Balla, Llorens Salvador & Delany, 2022*). Traditional ML Algorithms like SVMs and Decision Trees prove useful but require optimization for Arabic's unique features (*Alkhurayyif & Sait, 2023*). Neural networks effectively capture syntactic and semantic information but may overlook Arabic's rich morphological details (*Alkhurayyif & Sait, 2023*; *Alsubhi, Jamal & Alhothali, 2022*). Innovations like those by *Soumia (2024)* in developing temporal resources for Arabic, along with retrieval techniques like BM25 and Dense Passage Retrieval, stress the importance of tailoring tools and methods to the language's specifics. Custom tools like MADAMIRA are vital yet pose challenges in their creation and maintenance, highlighting areas for further research and development. Pre-trained models, such as BERT, have considerably increased Arabic NLP by handling complicated morphology. However, as demonstrated by *Alsubhi, Jamal & Alhothali (2021)*, the performance of these models varies significantly between datasets. The major gap is the lack of integration between traditional retrieval models, such as BM25, and neural-based retrieval systems. Combining the interpretability of classical models with the efficiency of neural networks may result in more consistent findings across datasets with diverse linguistic properties.

Discussion of RQ3: overcoming morphological challenges in Arabic

What tools and frameworks are utilized in this study for developing Arabic QASs, and how do these tools impact the system's performance? The study of tools and frameworks in AQAS development reveals substantial progress in solving Arabic language difficulties, with reliance on transformer-based models and NLP frameworks boosting system efficiency and accuracy (*Alami et al., 2023*; *Almiman, Osman & Torki, 2020*; *El Adlouni et al., 2021*). Camel Tools, ELMo vectorization, and Haystack NLP are useful for navigating Arabic syntax and semantics, but they may lead to technological lock-in, limiting innovation (*Alkhurayyif & Sait, 2023*; *Balla, Llorens Salvador & Delany, 2022*). Tailored techniques and novel methodology, such as those described in research by *Alsubhi, Jamal & Alhothali (2022)*, *Bakari, Bellot & Neji (2018)*, and *Mtibaa et al. (2018)*, are critical for solving specific language concerns, although they may have scaling issues. While beneficial, performance measurements should include user experience and flexibility in a variety of languages. The studies underline the need of interdisciplinary collaboration and open innovation in creating a dynamic and responsive AQAS development ecosystem. Despite these advances, Arabic's morphological variety remains a bottleneck. *Alwaneen et al. (2022)* highlight the limitations of existing techniques in dealing with dialectal variances, which are exacerbated by a lack of annotated datasets. Compared to tools such as Camel Tools and ELMo, hybrid AI approaches, as outlined in *Mahdi (2021)*, offer intriguing possibilities to overcome these difficulties by merging rule based and DL methodologies. Table 11 summarizes the tools, frameworks, and platforms used in the examined AQAS research, emphasizing their impact on system performance. The comparison indicates numerous key tendencies. Transformer-based models and frameworks, such as BERT, ON-LSTM, and Siamese architectures, regularly outperform classic machine learning approaches, especially when dealing with morphologically complicated Arabic questions. Specialized tools such as Camel Tools, MADAMIRA, and Haystack NLP are useful for syntactic and semantic processing, but they may provide issues due to technical lock-in and computational resources. Hybrid systems that combine neural networks with rule-based or classical retrieval methods have promise for balancing efficiency, interpretability, and accuracy, particularly when dealing with dialectal variances or domain-specific information. Finally, Table 11 highlights shortcomings in scalability and generalizability, underlining the importance of adaptive frameworks that can retain high performance across varied datasets. Overall, this table serves as a clear standard, directing tool selection in future AQAS development and suggesting areas for methodological innovation.

Discussion of RQ4: characteristics of Arabic QA datasets and their impact on system effectiveness

What are the specific contributions of AI, ML, and DL models in enhancing the accuracy and scalability of AQAS, and how do they address language-specific challenges such as Arabic morphology and dialectal diversity? AI, ML, and DL significantly improve AQAS's ability to evaluate and understand Arabic inquiries. Studies highlight the employment of BERT models, transformers, and new neural networks such as ON-LSTM and CNNs to handle Arabic syntactic and morphological difficulties, hence enhancing system performance (*Alami et al., 2023*; *Alsubhi, Jamal & Alhothali, 2022*; *Balla, Llorens Salvador & Delany, 2022*). While these technologies improve document retrieval and comprehension, they nevertheless demand a significant quantity of training data and computational resources, raising concerns about their applicability in resource-constrained scenarios (*Alkhurayyif & Sait, 2023*). Furthermore, studies show the importance of AI in semantic comprehension, emphasizing the need for a unified framework to integrate sophisticated AI methods into AQAS research (*El Adlouni et al., 2021*; *Othman, Alkhurayyif & Sait, 2022*; *Bakari, Bellot & Neji, 2018*). While datasets such as Arabic-SQuAD and TyDiQA-GoldP improve system robustness, their inherent flaws, such as translation problems and dataset sparsity, hamper AQAS development. *Alsubhi, Jamal & Alhothali (2021)* show that dataset quality has a direct impact on model performance, highlighting the importance of linguistic diversity in future datasets. Furthermore, combining domain-specific knowledge with broader resources could aid in understanding Arabic's complicated morphology and context-specific nuances.

To provide methodological clarity, we underline that the conclusions in this section are based on a structured literature review, rather than actual experimental research. The methodology builds on prior empirical studies to provide a thorough assessment of AI, ML, and DL contributions to AQAS. This approach is consistent with systematic review principles, allowing for critical examination and meaningful synthesis of existing literature to meet specific research objectives. Using these models, the discussion provides actionable ideas on enhancing the accuracy, scalability, and linguistic adaptability of AQAS systems.

Discussion of RQ5: attributes of Arabic QA datasets and their influence on system performance

How do the domain, size, and linguistic diversity of the Arabic QA datasets influence the system's performance in this study? The analysis of twelve studies highlights the diverse AQAS datasets, including ARCD, TyDi QA, Arabic SQuAD, MLQA, and XQuAD, which enhance system robustness and generalizability (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*; *Alsubhi, Jamal & Alhothali, 2022*). However, this diversity poses challenges in maintaining consistent evaluation metrics. While studies on Modern Standard Arabic and dialectal variations showcase the complexity of processing Arabic nuances (*Balla, Llorens Salvador & Delany, 2022*; *El Adlouni et al., 2021*; *Romeo et al., 2019*), the effectiveness of AQAS in handling these variations is underexplored. The use of translated datasets addresses resource scarcity but risks losing semantic integrity (*Lahbari, El Alaoui & Zidani, 2018*; *Othman, Alkhurayyif & Sait, 2022*). A more critical examination of how AQAS handles ambiguity and varying dataset characteristics is needed to assess their impact on system effectiveness and user satisfaction.

Discussion of RQ6: evaluation metrics and performance results of the QAS

What metrics are used to evaluate the QAS in this study, and what are the reported results in terms of accuracy, precision, recall, F1-score, or other metrics? The evaluation of AQAS uses a wide range of criteria that reflect the complexity of measuring these systems. To ensure clarity and minimize redundancy, we have combined the description of evaluation metrics into a single, comprehensive section.

Standard measures including as Accuracy, Precision, Recall, and F1-score have long been used to evaluate system performance, providing insights into the relevance and completeness of AQAS responses. However, recentstudies, such as

Alkhurayyif & Sait (2023), show that while high scores suggest great system performance, they also raise concerns about robustness when applied to various datasets. Furthermore, evaluation methods have evolved to include advanced assessment measures. Exact Match (EM), which evaluates rigorous answer correctness, and ranking-based metrics such as Mean Reciprocal Rank (MRR) and Precision@K have become popular for examining retrieval and ranking components in QAS. Such measurements are especially valuable in measuring how well a system ranks relevant replies, which addresses important aspects of user satisfaction and retrieval performance.

The variety of evaluation measures among studies highlights the necessity for a standardized evaluation approach for Arabic QAS. Without rigorous benchmarking, comparisons between research become inconsistent, making it impossible to draw firm conclusions about system strengths and limits.

The methodology used in this work synthesizes existing literature to provide a structured review of evaluation approaches in Arabic QASs, rather than presenting novel experimental results. This literature-based study assists in identifying gaps, establishing trends, and proposing modifications to match evaluation frameworks with current QA research standards.

By streamlining the discussion and deleting extraneous portions, we were able to provide a more structured and useful assessment of AQAS evaluation criteria while preserving methodological clarity.

Discussion of RQ7: identified research gaps, challenges, and proposed future directions

What research gaps or challenges does this study identify, and what specific future research directions does it propose? The study focuses on both technical infrastructure and linguistic issues in AQAS development, including advanced feature extraction and DL (*Alami et al., 2023*; *Alkhurayyif & Sait, 2023*; *Almiman, Osman & Torki, 2020*; *Alsubhi, Jamal & Alhothali, 2022*; *Othman, Alkhurayyif & Sait, 2022*; *Romeo et al., 2019*). However, scalability and resource constraints, particularly in underdeveloped countries, pose major challenges. Challenges such as morphological ambiguity and dialectal differences (*Alkhurayyif & Sait, 2023*; *Balla, Llorens Salvador & Delany, 2022*; *Soumia, 2024*; *Othman, Alkhurayyif & Sait, 2022*) necessitate flexible models, but the impact of socio-linguistic aspects requires further investigation. The paucity of annotated datasets (*El Adlouni et al., 2021*; *Lahbari, El Alaoui & Zidani, 2018*) emphasizes the importance of data enrichment and free access to resources. The emphasis is on semantic and logical analysis, including temporal processing (*Bakari, Bellot & Neji, 2018*; *Soumia, 2024*), which points to the need for sophisticated NLP tools, that potentially benefit from interdisciplinary approaches. A comparison of related works indicates significant gaps, including the absence of hybrid models to manage semantic and logical complexity. *Alkhurayyif & Sait (2023)* advocates for interdisciplinary techniques to augment datasets and improve algorithm resilience. Future research should emphasis scalability and flexibility in AQAS design, particularly for deployment in resource-constrained situations, as noted by *Mahdi (2021)*.

CONCLUSION AND FUTURE RESEARCH RECOMMENDATIONS

This study's in-depth examination of AQAS reveals substantial advances in the use of AI, ML and DL to improve Arabic language comprehension and processing. By rigorously examining twelve independent investigations completed between 2018 and 2023; this article demonstrates the critical contributions of sophisticated computational tools to AQAS. It also outlines the obstacles that researchers and developers confront, as well as suggestions for future study directions. This study is especially useful for academics, practitioners, and developers in the fields of AI and NLP who are working to improve AQAS technology. It gives these stakeholders a complete understanding of current approaches, identifies important obstacles, and proposes practical options for future developments, laying the groundwork for ongoing innovation and development in AQAS.

Main gaps identified

The analysis of AQAS identifies critical limitations, including a lack of high-quality, diversified annotated datasets that reflect Arabic's linguistic diversity and dialectal variances, which limit model efficacy. Hybrid AI models for increased semantic analysis are yet underexplored, limiting deeper language understanding. Furthermore, cutting-edge models like BERT suffer substantial computational demands, which limit scalability and usefulness, especially in resource constrained contexts. Arabic's rich morphology and dialectal variety highlight the need for sophisticated techniques to increase accuracy and performance while minimizing resource requirements.

Main challenges of AQAS

The main challenges confronting AQAS, as highlighted through the review, are multifaceted, encompassing both technical and linguistic hurdles. Linguistic diversity and complexity of the Arabic language, including its extensive morphology, dialectal diversity, and semantic nuances, pose significant challenges for AQAS development. Furthermore, despite advancements in current technological approaches, they frequently encounter technical limitations that impede their ability to fully meet the nuanced requirements of Arabic language processing. These limitations are particularly evident in terms of computational demand and the capability to adapt to the unique linguistic features of Arabic. However, the progress in the field of AQAS is significantly hindered by the limited availability of high-quality, diverse, and domain-specific datasets. These resources are crucial for the effective training and testing of AQAS models, and their scarcity not only represents a notable obstacle to advancements within this domain but also limits the potential to fully utilize AI capabilities in AQAS, thereby restricting the exploration and implementation of cutting-edge AI technologies to their fullest extent.

RQ-finding evidence summary

To make the claim-to-evidence chain clear, the primary conclusions are immediately linked to supporting evidence. For RQ1 (applied AI/ML methodologies), the findings suggest that sophisticated DL models such as CNNs, LSTMs, and BERT outperform classical ML methods in handling Arabic semantics, as illustrated in Table 11 and studies S3, S7, and S9. For RQ2 (major issues in AQAS), the review reveals that dialectal diversity, dataset scarcity, and computational efficiency remain persistent challenges, as evidenced by the study selection shown in Fig. 2 and articles S2, S5, and S10. For RQ3 (future research paths), the evidence indicates toward producing varied annotated datasets, adopting lightweight architectures, and investigating fairness-aware and hybrid models, as noted in the Future Research Recommendations and studies S1, S4, and S8. This mapping strengthens the reliability of the conclusions by linking each major takeaway to specific evidence within the review.

Future research recommendations

Future research on AQAS should prioritize the creation of varied, annotated datasets that account for Arabic's linguistic diversity, such as dialects and terminology. Exploring transfer and few-shot learning is critical for addressing the lack of high-quality Arabic data. Lightweight models, such as MobileNets and EfficientNets, should be created to reduce processing requirements while preserving accuracy. Hybrid AI models, such as merging Graph Neural Networks with RNNs or LSTMs, can help with semantic analysis by employing tactics like Semantic Role Labeling and Entity Recognition to improve query interpretation. Furthermore, tackling bias and fairness in AQAS models using Fairness Aware Learning Algorithms is critical for ensuring equitable outcomes across Arabic dialects and sociolinguistic groups.

Limitations of the study

This study has limitations. First, the emphasis on twelve selected studies may not accurately reflect the scope of AQAS research. Second, discrepancies in datasets, tasks, and evaluation criteria hinder direct comparability among studies. Third, the defined recency frame (2018–2023) prioritizes the most recent academic contributions while potentially ignoring earlier fundamental work or ongoing industrial advancements. Finally, the processing costs and contextual comprehension limits of present models limit their real-world utility. Despite these limitations, the study provides a rigorous synthesis of AQAS research, identifies important challenges, and makes actionable recommendations for future improvements in the field.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Azza Mohamed conceived and designed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.

- Khaled Abdelqader conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Khaled Shaalan analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data was not generated in this literature review.

REFERENCES

Abdallah A, Kasem M, Abdalla M, Mahmoud M, Elkasaby M, Elbendary Y, Jatowt A. 2024. ArabicaQA: a comprehensive dataset for Arabic question answering. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2049–2059.

Abdelqader KJ, Mohamed A, Shaalan K. 2023. Systematic review of automatic Arabic text summarization techniques. In: *Smart Innovation, Systems and Technologies.* Vol. 358. Berlin, Germany: Springer.

Al Chalabi HM. 2015. Question processing for Arabic question answering system. British University in Dubai (BUiD).

Al-Saleh AB, Menai MEB. 2016. Automatic Arabic text summarization: a survey. *Artificial Intelligence Review* **45(2)**:203–234 DOI 10.1007/s10462-015-9442-x.

Alami H, El Mahdaouy A, Benlahbib A, En-Nahnahi N, Berrada I, Ouatik SEA. 2023. DAQAS: deep Arabic question answering system based on duplicate question detection and machine reading comprehension. *Journal of King Saud University–Computer and Information Sciences* **35(8)**:101709 DOI 10.1016/j.jksuci.2023.101709.

Alanezi F, Alenezi M, Ghaith FB. 2023. Challenges in Arabic NLP and proposed solutions. *Int. J. Comput. NLP (IJCNLP).*

Alkhurayyif Y, Sait ARW. 2023. Developing an open domain Arabic question answering system using a deep learning technique. *IEEE Access* **11**:69131–69143 DOI 10.1109/ACCESS.2023.3292190.

Alkhurayyif Y, Wahab Sait AR. 2023. A comprehensive survey of techniques for developing an Arabic question answering system. *PeerJ Computer Science* **9(1)**:e1413 DOI 10.7717/peerj-cs.1413.

Almiman A, Osman N, Torki M. 2020. Deep neural network approach for Arabic community question answering. *Alexandria Engineering Journal* **59(6)**:4427–4434 DOI 10.1016/j.aej.2020.07.048.

Alsubhi K, Jamal A, Alhothali A. 2021. Pre-trained transformer-based approach for Arabic question answering: a comparative study. ArXiv DOI 10.48550/arXiv.2111.05671.

Alsubhi K, Jamal A, Alhothali A. 2022. Deep learning-based approach for Arabic open domain question answering. *PeerJ Computer Science* **8(1)**:e952 DOI 10.7717/peerj-cs.952.

Alwaneen TH, Azmi AM, Aboalsamh HA, Cambria E, Hussain A. 2022. Arabic question answering system: a survey. *Artificial Intelligence Review* **55(1)**:207–253 DOI 10.1007/s10462-021-10031-1.

Bakari W, Bellot P, Neji M. 2018. Towards an automatic text comprehension for Arabic question answering: semantic and logical representation of texts. In: Politzer-Ahles S, Hsu YY, Huang CR, Yao Y , eds. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Pennsylvania, US: Association for Computational Linguistics, 6–15.

Balla HAMN, Llorens Salvador M, Delany SJ. 2022. Arabic medical community question answering using ON-LSTM and CNN. In: *Proceedings of the 14th International Conference on Machine Learning and Computing (ICMLC 2022)*, 298–307 DOI 10.1145/3529836.3529913.

Bhoir V, Potey MA. 2014. Question answering system: a heuristic approach. In: *Proceedings of Fifth International Conference Applications of Digital Information and Web Technologies (ICADIWT)*, 1653-170 DOI 10.1109/ICADIWT.2014.6814704.

Bouziane C, Caan MWA, Tamminga HGH, Schrantee A, Bottelier MA, de Ruiter MB, Kooij SJJ, Reneman L. 2018. ADHD and maturation of brain white matter: a DTI study in medication-naive children and adults. *NeuroImage: Clinical* **17**:53–59 DOI 10.1016/j.nicl.2017.09.026.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems 33*, 1877–1901.

Calijorne Soares MA, Parreiras FS. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences* **32(6)**:6353–6646 DOI 10.1016/j.jksuci.2018.08.005.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12**:2493–2537 DOI 10.48550/arXiv.1103.0398.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the ACL*.

Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the NAACL-HLT*.

El Adlouni Y, Nahnahi NE, El Alaoui SO, Meknassi M, Rodríguez H, Alami N. 2021. Arabic Biomedical community question answering based on contextualized embeddings. *International Journal of Intelligent Information Technologies* **17(3)**:1–31 DOI 10.4018/IJIIT.2021070102.

El-Deeb R, Al-Zoghby AM, Elmougy S. 2018. Multi-corpus-based model for measuring the semantic relatedness in short texts (SRST). *Arabian Journal for Science and Engineering* **43(12)**:7933–7943 DOI 10.1007/s13369-018-3232-0.

Elsaid A, Mohammed A, Ibrahim LF, Sakre MM. 2022. A comprehensive review of Arabic text summarization. *IEEE Access* **10**:38012–38030 DOI 10.1109/ACCESS.2022.3163292.

Essam M, Deif MA, Elgohary R. 2024. Deciphering Arabic question: a dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways. *Artificial Intelligence Review* **57(9)**:251 DOI 10.1007/s10462-024-10880-6.

Hicke Y, Agarwal A, Ma Q, Denny P. 2023. AI-TA: towards an intelligent question-answer teaching assistant using open-source LLMs. ArXiv DOI 10.48550/arXiv.2311.02775.

Ishwari KSD, Aneeze AKRR, Sudheesan S, Karunaratne HJDA, Nugaliyadde A, Mallawarrachchi Y. 2019. Advances in natural language question answering: a review. ArXiv DOI 10.48550/arXiv.1904.05276.

Lahbari I, El Alaoui SO, Zidani KA. 2018. Toward a new Arabic question answering system. *The International Arab Journal of Information Technology (IAJIT)* **15(3A)**:96–105 DOI 10.34028/iajit.

Li H. 2018. Deep learning for natural language processing: advantages and challenges. *National Science Review* **5(1)**:243–256 DOI 10.1093/nsr/nwx110.

Mahdi AF. 2021. Survey: using BERT model for Arabic question answering system. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **12(13)**:7233–7729.

Mahdi HS, Alfadda HA. 2021. Measuring students' use of Zoom application in language course based on the technology acceptance model (TAM). *Journal of Psycholinguistic Research* **50(2)**:345–357 DOI 10.1007/s10936-020-09752-1.

Malik N, Sharan A, Biswas P. 2013. Domain knowledge enriched framework for restricted domain question answering system. In: *2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 137 DOI 10.1109/ICCIC.2013.6724163.

Mervin R. 2013. An overview of question answering system. *International Journal of Engineering Research and Advanced (IJRATE)* **1**:113–114.

Mtibaa A, Tourani R, Misra S, Burke J, Zhang L. 2018. Towards edge computing over named data networking. In: *2018 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 1–8 DOI 10.1109/EDGE.2018.00023.

Obeidat R, Heaton LJ, Tranby EP, O'Malley J, Timothé P. 2024. Social determinants of health linked with oral health in a representative sample of U.S. adults. *BMC Oral Health* **24(1)**:1518 DOI 10.1186/s12903-024-05257-8.

Othman S, Alkhurayyif M, Sait W. 2022. Advances in Arabic question answering systems. In: *ACM SIGIR*.

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**:1–67 DOI 10.48550/arXiv.1910.10683.

Romeo S, Da San Martino G, Belinkov Y, Barrón-Cedeño A, Eldesouki M, Darwish K, Mubarak H, Glass J, Moschitti A. 2019. Language processing and learning models for community question answering in Arabic. *Information Processing & Management* **56(2)**:2743290 DOI 10.1016/j.ipm.2017.07.003.

Sarkar A, Gupta R, Singh P. 2023. *Advancements in missing data imputation techniques: a comprehensive review*. Cham: Springer Nature.

Sequeda J, Allemang D, Jacob B. 2023. A benchmark to understand the role of knowledge graphs on large language models accuracy for question answering on enterprise SQL databases. ArXiv DOI 10.48550/arXiv.2311.07509.

Soumia Z. 2024. Exploring communication preferences for SDGs awareness among Algerian university students. SSRN DOI 10.2139/ssrn.5032449.

Sultana T, Badugu S. 2020. A review on different question answering system approaches. In: *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, 5793-586 DOI 10.1007/978-3-030-24318-0_67.

UNESCO. 2021. *World languages report: Arabic language facts and figures*. Paris: UNESCO Publishing.

Wang T, Zilinskas R, Li Y, Qu Y. 2022. Missing data imputation for a multivariate outcome of mixed variable types. ArXiv DOI 10.48550/arXiv.2206.01873.

United Nations. 2020. Official languages of the United Nations. *Available at https://www.un.org/en/our-work/official-languages*.

Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. 2021. mT5: a massively multilingual pre-trained text-to-text transformer. In: *Proceedings of the ACL*.

Zaki M. 2019. Digital transformation: harnessing digital technologies for the next generation of services. *Journal of Services Marketing* **33(6)**:751–762 DOI 10.1108/JSM-01-2019-0034.