Artificial Intelligence Review (2025) 58:127
https://doi.org/10.1007/s10462-025-11122-z

A review on persian question answering systems: from traditional to modern approaches

Safoura Aghadavoud Jolfaei[1] · **Azadeh Mohebi**[1]

Accepted: 23 January 2025 / Published online: 13 February 2025

Abstract
Question answering systems (QAS) are designed to answer questions in natural language. The objective of these types of systems is to reduce the user's effort to manually check the retrieved documents to find the answer to the query in natural language and to create an accurate answer to the user's query. In recent years, with the emergence of Large Language Models (LLMs), these systems have evolved significantly across different languages. However, the development of QAS in low resource languages such as Persian, while progressing, still faces unique challenges. Development of these systems has become problematic in Persian language due to the lack of comprehensive processing tools, limited question answering datasets, and specific challenges of this language. The current study provides a brief explanation of these systems' evolution from traditional architectures to LLM-based approaches, their classification, the challenges specific to Persian language, existing question-answering datasets and language models, and studies conducted concerning Persian QAS.

Keywords Persian question answering system · Information retrieval · Natural language processing · Large language models

1 Introduction

Due to the rapid growth of information on the web, access to information and management of available big data has become crucial. The growing demand for access to different types of information available online has led researchers' interest to retrieve information in a wide range that is related to areas such as question answering, topic discovery and tracking, summarizing, multimedia retrieval, text mining, and text structuring. Varieties of methods have been applied in artificial intelligence, particularly natural language processing, to

✉ Azadeh Mohebi
mohebi@irandoc.ac.ir

Safoura Aghadavoud Jolfaei
aghadavood@students.irandoc.ac.ir

[1] Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran

design tools and methods. Natural language processing is one of the progressing fields in artificial intelligence, with the possibility of developing automatic methods of answering users' questions. The ability to read text and answer questions has become one of the main streams in the search for understanding natural language. In 2011, IBM presented its deep question answering system called Watson, which served as a turning point in the development of QAS, as a result of which natural language processing experienced unexpected progress (Ferrucci et al. 2010). Answers to questions are extracted in natural language by QAS. These systems provide the possibility to determine answers to a certain question, without browsing search results manually (Elfadil et al. 2021). QAS is considered an area for developing search engines to provide the user with the precise answer rather than an ordered list of related documents (Dimitrakis et al. 2019). Comparing QAS with search engines, it can be claimed that effective complex questions and answers can help improve search systems. When a user searches for information, search engines usually offer a list of documents, each of which the user must select and check to find the desired information. In addition, available search systems do not show the level of user satisfaction with the search result via which the search policy can be improved (Chali et al. 2015), while in QAS, the answer is provided according to the user's question and it is possible to evaluate the user's level of satisfaction with the given answer and the performance of the system based on the user's feedback.

The landscape of question answering systems has evolved dramatically since these early developments. While traditional QAS focused on information retrieval and document processing, the advent of LLMs has transformed how these systems approach natural language questions. Despite these advances, languages with limited resources face unique challenges.

Currently, various QA approaches have been developed, from traditional information retrieval-based systems (Dimitrakis et al. 2019; Jurafsky and Martin 2020) to neural approaches (Rogers et al. 2020) and modern LLM-based solutions(Anil et al. 2023; Achiam et al. 2023). This evolution has transformed how systems approach natural language questions, moving from explicit information retrieval to more sophisticated language understanding and generation capabilities.

The morphological complexity of Persian can make the development of QASs even more complicated. In Persian, prefixes and suffixes are very common, which causes retrieving relevant documents a challenging task. Information retrieval is an essential step in the development of QAS. The most related documents to a given question need to be retrieved and processed to form an answer. Finding the most related document depends on finding the keywords of the question in the document.

In this research, previous researches in Persian QAS is reviewed with a focus on the following issues:

- Identifying QAS developed in Persian,
- Reviewing the studies conducted on the development and design of modules used in Persian QAS,
- Examining the available tools used by the researchers for Persian QAS development,
- Investigating available datasets in Persian for QAS,
- Examining the criteria used to evaluate QAS,
- Identifying challenges and trends for Persian QAS research.

Our study focuses on the development of QAS in the Persian language, examining both traditional approaches and recent LLM innovations. while studies have been conducted on such systems in English, resulting in a great number of publications (Bouziane et al. 2015; Mishra and Jain 2016; Saini and Yadav 2017; Zaib et al. 2021). However, fewer studies have been conducted in other languages (Alwaneen et al. 2022).

To address this, we conduct a comprehensive review spanning traditional methods, neural approaches, and recent LLM-based systems, examining datasets, methods, and language models used in Persian QAS research.

In this work, we strive to make a thorough investigation into the challenges that may arise when developing QAS for Persian. It is a fact that Persian, being a language with limited resources, poses unique obstacles because there are limited training data and linguistic resources available for the language. QAS development is further complicated by the complexity of the Persian text, especially in terms of its structure and grammar. As a result of identifying these challenges and addressing them, we are paving the way for future improvements in the research of Persian QAS in the future.

Furthermore, our study emphasizes the importance of benchmark datasets specifically tailored to Persian, since there is a lack of standardized datasets that hinders fair comparisons between existing systems. A comprehensive natural language processing tool that takes Persian language intricacies into account is also needed. By leveraging deep learning techniques, we can significantly improve Persian QAS performance, provided we overcome limited datasets. Ultimately, our research contributes to QAS progress in Persian, opening doors for further exploration and development in this field.

This review on Persian QAS provides valuable implications for other low-resource languages. It raises awareness, offers guidance, promotes resource and tool adaptation, encourages collaboration, and contributes to the advancement of language technology in the context of question answering systems for languages with limited resources.

In the next section, the definition, classification, and architecture of these systems are briefly stated, as well as the criteria generally used for evaluation. Then, the challenges that usually exist in the Persian language for developing QASs are discussed. Next, the question and answering systems developed in Persian, systems developed concerning downstream tasks in these systems, and available tools for the development of these systems in Persian are reviewed. Finally, the conclusion and future trends of these systems are investigated.

2 Question and answering systems

In this section, the main concepts of QAS including different types and classifications, architectures, and evaluation criteria are discussed.

2.1 Classification criteria

Question answering systems have been traditionally classified along multiple dimensions, as identified by various researchers. The primary classification criteria include:

- Data source-based classification (Bouziane et al. 2015; Jurafsky and Martin 2020).
- Processing technique-based division (Huang et al. 2019; Yogish et al., 2016).

- Domain-based categorization into Open, Closed, and Restricted domains (Mishra and Saxena 2019; Pundge 2016).
- Answer-type classification (Franco et al. 2020).
- Question-type classification (A. Mishra and Jain 2016).

Based on the previous studies, Table 1 shows the classification criteria and the types of them.

While Table 1 represents the traditional classification of question answering systems, the advent of LLMs has fundamentally transformed these classification criteria themselves. Prior to LLMs, these classifications were essential because each category demanded different processing approaches. For example, factoid questions needed named entity recognition(Lopez et al. 2013), while hypothetical questions required information retrieval methods (Mishra and Jain 2016).

Table 2 shows that Classification criteria, Traditional Limitation which highlights the historical constraints that necessitated strict categorization in conventional QA systems, the LLM-Based Evolution which demonstrates how modern LLM-based approaches have overcome these limitations, introducing unified processing and flexible boundaries across all criteria and the Supporting Evidence column which cites recent research that validates these transformations, providing empirical evidence for each advancement. This evolution reflects a broader shift from rigid, categorization-dependent processing to a more flexible, unified approach where traditional classifications serve more as analytical frameworks than architectural necessities (Zan et al. 2023).

The LLM Unified Approach represents a fundamental shift in how question answering systems handle traditional classification criteria. This transformation is motivated by several key factors. Figure 1 shows the LLM transformation in question answering systems.

These classifications, while still useful for evaluation and analysis, no longer represent strict technical boundaries in system implementation. Instead, they serve more as conceptual frameworks for understanding QAS capabilities and performance characteristics.

Despite the transformative capabilities of LLMs in question answering systems, several significant challenges still remain. While Retrieval-Augmented Generation (RAG) has emerged as a solution to reduce hallucination by providing contextual grounding through domain-specific document retrieval (Shuster et al. 2021), its effectiveness varies considerably across different domains. General domain retrievers often underperform compared to domain-specific fine-tuned models (Gao et al. 2023). Additional challenges include the

Table 1 Traditional question answering systems criteria

Criteria	The types of criteria
Type of question	Factoid (Calijorne Soares and Parreiras 2020; Heie et al. 2012), List (Calijorne Soares and Parreiras 2020; Heie et al. 2012), Hypothetical (Ojokoh and Adebisi 2019), Confirmation (Ojokoh and Adebisi 2019), Definition or Casual (Dimitrakis et al. 2019; Neves and Leser 2015; Ojokoh and Adebisi 2019) and Opinion (A. Mishra and Jain 2016)
Domain	General, Restricted (Antoniou and Bassiliades 2022; Latifi 2018)
Type of answers	Extracted, Generated (Dimitrakis et al. 2019; Mishra and Saxena 2019)
Data sources	Raw Text, Knowledge Base (Jurafsky and Martin 2020) and Hybrid (Dimitrakis et al. 2019)

Table 2 Evolution of QA classification: from traditional to LLM-based approaches

Classification criteria	Traditional limitation	LLM-based evolution	Supporting Evidence
Question type classification	Rigid categorization requiring separate processing for each type	Unified processing across all question types without explicit categorization	(Lan et al. 2023; Li et al. 2024; Prabhu and Anand 2024; Wang et al. 2023a, b)
Domain classification	Strict boundaries between general and specialized domains	Dynamic adaptation across domains	(Liu et al. 2023; Wang et al. 2024; Wang et al. 2023a, b; Zhang et al. 2023)
Data source classification	Separate pipelines for different source types	Unified processing of multiple sources through Retrieval-Augmented Generation	(Gao et al. 2023; Wang et al. 2024; Wang et al. 2023a, b)
Answer type classification	Binary distinction between extraction and generation	Hybrid approach combining both methods with verification	(Wang et al. 2023a, b)

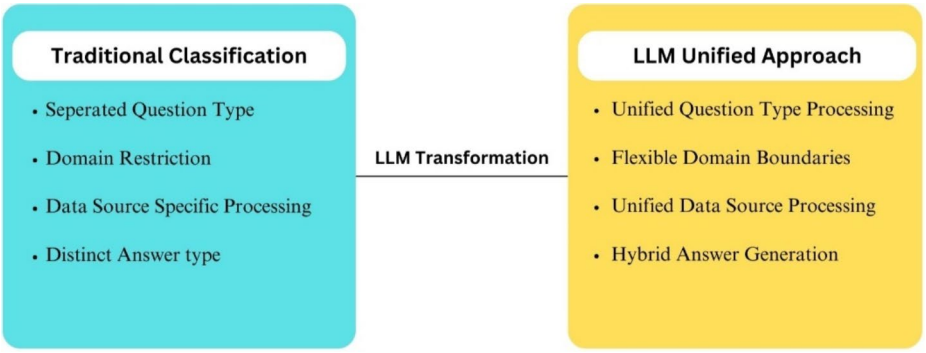

Fig. 1 From traditional classification question answering system to LLM-question answering systems

computational cost of maintaining and running large models, the need for extensive prompt engineering, and potential inconsistencies in response quality. These limitations become particularly apparent in specialized domains where traditional QA systems with domain-specific training still maintain advantages in accuracy and reliability(Zhao et al. 2022). Furthermore, LLMs face challenges in verifiable fact-checking, explicit reasoning transparency, and maintaining consistent performance across different question types (Thakur et al. 2021). One promising approach to address these limitations is the integration of knowledge graphs with LLMs, where structured knowledge representations can help anchor responses to verified facts, reducing hallucination and improving answer precision (Bui et al. 2024;

Zhao et al. 2024). As the field evolves, addressing these limitations while preserving the benefits of unified processing remains a key research challenge.

2.2 QAS architecture

The architecture of QAS has traditionally been shaped by the nature of their underlying knowledge sources (Dimitrakis et al. 2020). For text-based sources, systems employ information retrieval methods combined with machine comprehension to locate and extract answers. When working with structured data in formats like Resource Description Framework (RDF), systems utilize graph processing methods and SPARQL queries(Antoniou and Bassiliades 2022). In cases where both structured and unstructured sources are available, or when single-source approaches prove insufficient, hybrid methods are implemented to leverage multiple knowledge sources effectively (Dimitrakis et al. 2020).

With the advent of LLMs, these architectural paradigms have evolved significantly while maintaining their core relationship with knowledge sources. Figure 2 illustrates the evolution of QAS architectures, from traditional approach to modern LLM-based systems.

Text-based Architecture (Fig. 2a) processes questions through a sequential pipeline, starting with question analysis and query reformulation, followed by document retrieval and ranking, and concluding with answer extraction from relevant passages. This traditional approach leverages information retrieval techniques and machine comprehension to find answers in unstructured text (Ojokoh and Adebisi 2019).

Graph-based Architecture (Fig. 2b) transforms natural language questions into structured queries, primarily using SPARQL for knowledge graph traversal. Through query processing, graph navigation, and pattern matching, it extracts precise answer from structured knowledge sources like RDF graphs. This approach provides explicit reasoning paths and structured answer generation(Antoniou and Bassiliades 2022; Dimitrakis et al. 2020).

LLM-based Architecture (Fig. 2c) combines input processing, core LLM functions, and answer verification, while addressing the critical challenge of hallucination through external knowledge integration.

Knowledge graphs help LLMs break down complex questions into simpler parts using logical reasoning chains (chain-of-thoughts)(Wei et al. 2022).By using RAG (Gao et al., 2023) and knowledge graphs as external knowledge sources, the system ensures LLM responses are based on verified facts rather than potentially incorrect generated informa-

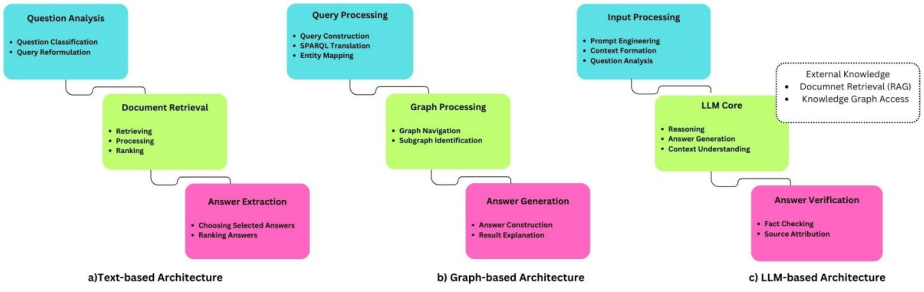

Fig. 2 QAS architectures: (a) Traditional text-based QAS with question analysis and information retrieval pipeline, (b) Graph-based QAS utilizing semantic knowledge processing, and (c) LLM-based QAS integrating external knowledge sources to enhance factual accuracy

tion, making answers more accurate and trustworthy (Wang et al. 2023a, b). This hybrid approach maintains the powerful reasoning capabilities of LLMs while ensuring factual consistency through structured and unstructured knowledge sources(Daull et al. 2023; Li et al. 2023).

2.3 QAS evaluation criteria

Evaluating QAS is inherently complex due to the diverse nature of questions and possible answers. While traditional evaluation methods using labeled data and measurable scores remain fundamental (Farea et al. 2022), the emergence of LLM has introduced new evaluation challenges and metrics.

2.3.1 Traditional evaluation methods

Classic datasets.

Datasets to evaluate QAS are different in format, type, and number of questions and relevant answers. Each question may have one or none or several answers. When there is no relevant answer to the query, the answer is usually returned by the annotators with the label "none", as the golden answer during the evaluation process. There are several golden standard datasets in this field (Franco et al. 2020) such as QALD (Cimiano et al. 2013), WebQuestions (Berant et al. 2013), SQuAD (Rajpurkar et al. 2016), and SimpleQuestions (Bordes et al. 2015).

Standard metrics.

The evaluation process requires comparing system outputs against reference answers (golden answers) using various metrics (Farea et al. 2022). The most widely used metrics are (Chen et al. 2019):

- **BLEU** (Papineni et al. 2002), the idea is that a good translation should produce a text like a human-translated text. To do this, Measures n-gram overlap with reference text.
- **METEOR** (Banerjee & Lavie, 2005), Enhanced BLEU with stem and synonym matching.
- **ROUGE** (Lin 2004), Evaluates summary quality through n-gram overlap.
- **Sentence Mover's Similarity (SMS)** (Clark et al. 2019), Compares sentence vector similarities.
- **BERTScore** (Zhang et al. 2019), Uses BERT for semantic similarity.
- **Exact Match** (Rajpurkar et al. 2018), Binary score for perfect matches.
- **Precision**, Ratio of correct to total answers.
- **Recall**, Ratio of found correct answers to all possible correct answers.
- **F1** (Rajpurkar et al. 2018), It tries to find harmony between recall and precision.
- **MRR**, It evaluates the probability of correctness of each of the answers related to the sample query and ranks them.

2.3.2 Modern evaluation approaches

Factual Correctness metrics

Modern QAS, particularly LLMs, require rigorous evaluation of their factual accuracy to ensure reliable information delivery and prevent hallucination(Agarwal, n.d.; Tam 2023).

- Faithfulness Score(Fayyaz et al. 2024): Measures consistency between answer and knowledge sources.
- Hallucination Detection (Liu et al. 2019): Identifies generated content not supported by sources.
- Source Attribution(Do et al. 2024): Evaluates accuracy of cited information.

Reasoning Quality Metrics:

Today's complex questions require evaluating not just the final answer, but also how the system arrives at that answer through logical steps(Wolfson et al. 2023).

- Chain of Thought (CoT) evaluation assesses how well a QAS explains its reasoning process, by examining whether each logical step leads naturally to the next and ultimately to the correct answer(Fu et al., n.d.; Wei et al. 2022). This is particularly important for complex questions where the system must show its work, similar to how a human would solve a problem step by step.
- Multi-hop accuracy focuses on questions that require connecting multiple pieces of information to arrive at an answer, measuring the system's ability to navigate through different facts or documents to find and combine relevant information. For example, answering "Who was the first wife of the president who established NASA?" requires finding information about NASA's founding president and then finding information about his first wife(Khandelwal et al. 2024).
- The consistency score examines whether a system provides stable, reliable answers when the same question is asked in different ways or with different contexts, which is crucial for evaluating the robustness of LLM-based QAS (Saxena et al. 2024).

Together, these metrics help ensure that modern QAS not only provide correct answers but also arrive at them through sound reasoning processes.

3 Persian language

With over 110 million speakers in countries such as Iran, Afghanistan, and Tajikistan, Persian language has also significantly impressed the languages of the neighboring countries, such as Turkish, Armenian, Georgian, and Indian. Its alphabet consists of 32 letters written in a right-to-left order (Amirkhani et al. 2023). Moreover, Persian is among the top 25 languages in the world (Salemi et al. 2021).

Persian has certain features, which distinguish it from other well-studied languages. In terms of script, Persian is similar to Semitic languages (like Arabic), and in terms of linguistics, it is among Indo-European languages (Snell 2009), and can be claimed to be related

to most European languages, especially those of the northern Indian Peninsula. All these features make Persian a special case for research. Due to the lack of specialized language understanding criteria that can result in advancement in a wide range of tasks, performance measurability and development of natural language understanding models are limited in Persian.

In recent years, an increasing number of studies have been conducted on QAS in Persian language. QA tasks in Persian are of great significance because of the high volume of Persian information on the internet and increased users demand. However, limited research has been conducted in this area, and is still at initial levels, compared to the English language. Natural language processing systems that work well with other languages do not have satisfying performance in Persian due to the challenges posed by this language. In the following, the most important challenges are listed.

3.1 Challenges in Persian language

According to previous studies (Habib 2021; Shamsfard 2019), the challenges of the Persian language can be divided into two categories:

Challenges at the character level.

- Due to similarities in writing styles between Persian and Arabic, some letters in Persian can be written in different forms. Sometimes Arabic letters substitute Persian ones, such as "ی" and "ک" while the ASCII codes in these languages are different. The proposed solution is to use a normalizer with Persian texts to integrate different formats of letters and figures.
- Beginning words with a capital letter in languages such as English makes a great contribution to NER tasks, which is a great step in answering factoid questions. However, there is no capital letter as the initial letters of nouns/names, which makes NER operations ambiguous for question analysis and answer extraction. Hence, the system requires more tools to distinguish between nouns and other functional words such as adjectives or verbs.
- Words containing what is called "Hamza" in Persian, which can appear in the words in different forms (ء, أ, إ, ؤ, ئ, etc.); for example, "لئاسم" vs. "لیاسم" which means "problems".
- The role of a comma in conveying the meaning and message of sentences; for example, "تسا رامیب ، میرم ردپ" vs. "ردپ، میرم بیرا تسا" which means, "Mary's father is sick" vs. "Father! Mary is sick".
- Adding space in some words leads to a change in the meaning (e.g. "ردام" which means "mother" vs. "رد ام" which means "we in").
- Many multi-words terms exists in Persian that are written with half-space between the words, while using space instead of half-space changes the meaning of words and sentences. For instance, the word "رتنیگنس (means *heavier*) consists of two words: 1) "نیگنس" and, 2)"رت", with a half-space between the two. The first word means *heavy* and the second word means *wet*. If space is used instead of half-space, the meaning will be different.

Challenges at the word level.

- The presence of multi-unit tokens, that can be written with different spellings; this challenge can also be addressed by integrating the tokens via zero-with-non-joiner (ZWNJ) characters.
- The formats of plural words that originally come from Arabic and do not follow the ordinary rules of pluralization. The solution is to remove Arabic plural words.
- The existence of different forms for similar words (e.g., روتارپیما & روطارپیما), both mean "emperor" (Amirkhani et al., 2020).
- Different words for foreign terms and concepts such as "computer" and its Persian equivalent "رایانه" (pronounced as *Rayaneh*) which are both common.
- The presence of words with different meanings that have the same spelling. (e.g., "ریش" means "lion", vs. "ریش" means "milk")
- Words with both formal and informal forms.
- There are numerous multi token verbs in which each token may have a different meaning on its own, but when joined as a single verb, their meaning will be different. For instance the verb "تروص ترفتم تسا" which means "have been done", has three different tokens: "تسا", "ترفتم", and "تروص". The token "تسا" means "is" in English, the token "ترفتم" as an adjective means "solemn" and as a verb means "taken", and the token "تروص" as a noun means "face" while as a part of a verb means "done".
- Despite Arabic, French, and English, there is no difference between feminine and masculine words, and all have the same pronouns.

3.2 Processing tools in Persian

There are many tools for processing English Language, which are not suitable for processing Persian Language. Persian is written in a right-to-left order while English is written left to right. Part of speech (POS) is different in the two languages; also ambiguity in word morphology adds to the difficulty of Persian processing. The Arabic alphabet has four letters less than Persian, and although the two languages may look similar, they are completely different in writing. Moreover, Arabic is also faced with a lack of tools and resources. Hence, using tools from other languages is not a solution and makes the task even more complicated. Thus, language processing tools must be designed specifically for Persian (Habib 2021).

Certain tools have been developed for different processing tasks in Persian, which only support the desired task such as POS. Table 3 lists comprehensive tools which have already been developed for processing Persian (Etezadi et al. 2022). Each row shows the name of the preprocessing tool and each of the columns shows the characteristics of the tools (Table 3).

4 QA datasets, language models and LLMs in Persian

The evolution of language technology has been dramatic, from basic rule-based systems to neural networks, and now to LLMs. While these LLMs show remarkable abilities in many languages, they still struggle with resource-limited languages like Persian. Despite advances in multilingual LLMs and zero-shot learning, creating effective QAS still requires quality, domain-specific data to ensure reliable performance.

Table 3 Comprehensive tools for processing persian language

Tool	Ezafe[a]	Chunker	Constituency	Dependency	POS	Lemma	Normalizer
Parsivar[b]	×	✓	×	✓	✓	×	✓
Text_mining[c]	×	×	×	×	✓	✓	✓
DadmaTools[d]	✓	✓	✓	✓	✓	✓	✓
Hazm[e]	×	✓	×	✓	✓	✓	✓
Stanza (Qi et al. 2020)	×	×	×	✓	✓	✓	×

[a] Ezafe is a grammatical particle in Persian language that links two words together. Ezafe carries valuable information about the grammatical structure of sentences(Etezadi et al. 2022)

[b]ICTRC/Parsivar: A Language Processing Toolkit for Persian (github.com).

[c]Persian text mining tool: services for processing Persian texts (text-mining.ir).

[d]Dadmatech/DadmaTools: DadmaTools is a Persian NLP tool developed by Dadmatech Co. (github.com)

[e]Sobhe/hazm: Python library for digesting Persian text. (github.com)

Persian language is one of the languages that lack linguistic resources such as corpus labeling, QA datasets, pairs of paraphrases, chats, metaphors, etc. Although Persian language datasets are available for various NLP tasks such as sentiment analysis, natural language understanding, and machine translation, they are significantly smaller and less comprehensive compared to English datasets and other high-resource languages. This deficiency can be compensated by creating resources and tools for languages with few resources or by adapting to the tools and resources of other languages (Shamsfard 2019). While traditional resource limitations persist, new challenges have emerged. LLMs often struggle with Persian's complex morphology and dialectal variations, leading to potential hallucination and inaccurate translations(Alam et al. 2024).

4.1 QA datasets

There are four approaches to creating non-English datasets (Abadani et al. 2021):

1. Using native annotators: This approach is done using crowd workers and includes native language texts. The models trained on these data sets usually have higher quality than the ones trained on machine-generated datasets. However, this approach is expensive and requires extensive effort.
2. Using Neural Machine Translation (NMT) on existing English datasets: Translating the English dataset into the target language and fine-tuning the language model on the translated dataset. Recent research has shown that the data collected through translation artifacts is a poor method (Artetxe et al. 2020). Any natural language processing dataset needs to show the distribution of target language tokens and their relationship in the context. Therefore, it is better to avoid relying too much on the automatic conversion of languages with high resources to minimize unnatural samples or artifacts (Khvalchik and Galkin 2020).
3. Translation of existing datasets along with annotated datasets in native language: A combination of the two previous approaches.
4. Creating a cross-lingual evaluation standard: The goal is to develop a cross-lingual question and answering models. Unlike multilingual models, these models can be transferred to the target language without the need for training data in that language. XQuAD

(Artetxe et al. 2019) contains 1190 examples of SQuAD which has been translated by professionals into ten different languages. MLQA (Lewis et al. 2020) is a combination of 12,000 English examples in 6 other languages; however, Persian language is absent in both.

The first three approaches lead to the creation of new monolingual datasets, while the fourth approach will create multilingual datasets that will train multilingual models.

Recent developments in LLMs have introduced new possibilities for creating datasets in low-resource languages like Persian. LLMs can assist in dataset creation through multiple approaches: direct generation of question-answer pairs using few-shot prompting(Leite and Cardoso 2024; Schmidt et al. 2024) and enhancement of translation quality through cross-lingual verification. While these approaches show promise in reducing costs compared to human annotation (Chan et al. 2024), they still face challenges such as hallucination risks and the need for factual verification.

A hybrid approach combining LLM generation with human verification has emerged as a practical solution, offering better quality than pure machine translation while being more cost-effective than complete human annotation.

Current QA datasets in Persian are:

- **PeCoQ** (Etezadi and Shamsfard 2020b) is a QA dataset on knowledge graphs, which contains 10 thousand complex questions and answers extracted from the FarsBase knowledge graph (Sajadi et al. 2018). For each question, there is a SPARQL query and two paragraphs written by linguists. There are many complexities in this dataset, such as multi-relation, multi-entity, sequential and time constraints. FarsBase is a knowledge graph for the Persian language, whose data has been extracted from the Persian version of Wikipedia. It is a set of subject-relation-object triples. Three categories of complex questions are defined in this dataset: (1) multi-entity and multi-relation questions (2) sequential questions (3) temporal questions.
- **PerCQA** (Jamali et al. 2021) dataset includes questions and answers taken from the most famous Persian websites. Having collected the data, a detailed instruction has been prepared for annotation according to which QA pairs are created in the SemEvalCQA format. This dataset contains 989 questions and 21,915 answers. It is a Community Question Answering dataset that allows web users to achieve answers to their questions from experts in the field. The ability to automatically find relevant questions to reuse their existing answers (question retrieval) and search for relevant answers among a large number of answers for a given question (answer selection) is one of the most famous tasks in this collection. This dataset includes questions and answers posted by users from the forum of one of the most famous Persian language sites called “ninisite” (a website providing guidance on pregnancy and child upbringing). The focus of the article related to this dataset is on data collection and annotation.
- **ParSQuAD** (Abadani et al. 2021) is a Persian version of the SQuAD dataset (Rajpurkar et al. 2016) translated using Google Translate. Two versions of this dataset have been presented: ParSQuAD (manual) and ParSQuAD (automatic) with 25 thousand and 70 thousand samples. To create ParSQuAD (manual) after the automatic translation of SQuAD, some translation errors have been manually corrected. ParSQuAD (automatic) is the result of the automatic translation of a part of SQuAD without correction. The

structure of this dataset is similar to the structure of SQuAD, which has a tree structure. It includes several titles, each title is related to phrases, and each phrase has two types of questions: answerable and unanswerable. Each question has a set of answers with a number corresponding to the beginning of the answer domain.

- **PersianQuAD** (Kazemi et al. 2022) is a collection of questions and answers for the Persian language, prepared in four steps: (1) Selection of Wikipedia articles, (2) Creation of question-answer pairs, (3) Preparation of three candidates for the test set, and (4) Data quality monitoring. This dataset contains 20 thousand questions and answers created by native annotators on Persian Wikipedia articles. The answer to each question is part of the text of the corresponding article. Project Nayuki's Wikipedia's internal PageRanks were used to retrieve 10 thousand Persian articles from Wikipedia with the highest ranking. Finally, 26,417 paragraphs with different topics have been obtained. This dataset is divided into two subsets, train, and test, with 18,567 and 1,000 samples respectively, and the test dataset is randomly selected from the dataset. Three pre-trained language models have been implemented on this dataset, namely ParsBERT (Farahani et al. 2021), MBERT (Conneau et al. 2020), and ALBERTA-FA (Conneau et al. 2020).
- **PQuAD** (Darvishi et al. 2023) is a collection of reading comprehension data on Persian Wikipedia articles. It has 11 thousand phrases and 80 thousand questions with answers, each phrase contains 7 questions and 25% of the questions are unanswered. ParSQuAD and PersianQA[1] datasets are both automatically collected with different methodologies; however, they cover a small number of samples and, hence, are not suitable for supervised learning frameworks. The process of creating this dataset includes three steps: passage curation, question-answer pair annotation, and additional answer collection. In passage curation, the most important articles in Persian Wikipedia are selected based on two criteria: (1) having an info-box, and (2) Being among the top pages in Persian Wikipedia. With the PageRank algorithm, it retrieves 600,000 articles. A graph is made for these articles based on their links and the rank of collected articles using the PageRank method in the NetworkX library. Unanswered questions are shown with no answer, and the answer to the questions is shown by specifying the answer domain.
- **ParsSimpleQA** (Babaei Giglou et al. 2022) is a semi-automatically dataset which is created in two stages: Creating single-relational question patterns and creating simple questions and answers using templates, entities and relationships from Farsbase (Sajadi et al. 2018) automatically. Using deep learning and information retrieval techniques, they have developed a simple question-and-answer system that ensures the reliability of the dataset.
- **FarsQuAD** (Kazemi et al. 2023) a Persian QA dataset for the news domain. 500 news articles were randomly selected from Persian news websites for this dataset. They extracted 1000 paragraphs from the selected news articles and kept news articles with at least 500 characters. Using Persian news articles, 10 native Persian speakers created questions. The dataset has a total of 600 question-answer samples.
- **IslamicPCQA** (Ghafouri et al.,2023) is a dataset for answering complex question based on non-structured information sources. It consists of 12,282 question-answer pairs extracted from 9 Islamic encyclopedias. Among these data samples, 800 were randomly selected for analysis and study. Based on HotpotQA's English dataset approach (Yang et al. 2018), this dataset was customized for Persian's complexities. It takes more than one

[1] sajjjadayobi/PersianQA: Persian (Farsi) Question Answering Dataset (+Models) (github.com).

paragraph and reasoning to answer questions in this dataset. No prior knowledge base or ontology is required to answer the questions. The dataset has supporting facts and key sentences for exact reasoning. Using non-constructed Islamic resources in Persian, this research aims to generate complex multi-hop questions.

- **MeDiaPQA** (Medical Dialogue Persian QA) (Shahriar, 2023) is a specialized dataset focusing on medical domain question answering, containing 7,704 question-answer pairs extracted from 15,748 doctor-patient dialogues. The dataset's unique contribution is its coverage of more than 70 medical specialties, structured in a multiple-choice answer format.
- **FarsiQuAD** (ForutanRad et al. 2024) is a Persian question-answering dataset with two versions: a smaller one with 10,000 questions and a larger one with 145,000 questions. Created by human annotators using Persian Wikipedia articles, The dataset follows the same format as the popular SQuAD (Rajpurkar et al. 2016).

4.2 Language models and LLMs for Persian

In recent years, data-based neural models and machine learning problems have gained considerable success. In the field of natural language processing, transformers (Vaswani et al. 2017) and pre-trained language models such as BERT (Devlin et al. 2019) have resulted in great performance leaps in downstream tasks. The landscape has further evolved with the emergence of LLMs. The following sections discusses the related studies about traditional pre-trained models, and then the recent developments in Persian language affected by LLMs.

4.2.1 Traditional pre-trained models

- **XLM-RoBERTa** (Conneau et al. 2020) is a transformer-based language model. Like RoBERTa monolingual language model, it does not use Next Sentence Prediction (NSP) for training and only uses the multilingual Masked Language Model for training. This model can be trained in 10 different languages and has a powerful multilingual pre-trained language model.
- **MBERT** (Devlin et al. 2019), Google's deep bidirectional language model, which has been trained in 104 languages from Wikipedia, including Persian, and has shown good performance on NER, QA, and POS tasks.
- **ALBERT** (Lan et al. 2019) is a lighter version of MBERT with fewer parameters and faster training, which has outperformed MBERT in some tasks. ALBERT-FA is the Persian version of ALBERT which has been trained on Persian texts.
- **ARMAN** (Salemi et al. 2021) is a pre-trained transformer-based encoder-decoder model. Important sentences are selected based on the modified semantic score of the document to form a summary of it. For a more accurate summary and similar to human writing patterns, the sentences have been reordered. Six summarization tasks of Persian language were used for evaluation. Experimental results showed satisfying performance of the proposed model on all the six downstream summarization tasks measured by ROUGE and BERTScore. Finally, human evaluation was carried out and showed that the use of semantic score significantly improved the summarization results. The focus of

this article is to summarize the Persian text. However, the presented method is language independent. First, semantic similarity scores are used in sentence selection to create a pseudo-summary of the document. In short, selecting important sentences based on semantic scores in a self-supervised way creates a summary for the relevant document in the dataset. Models were fine-tuned on six downstream tasks. In addition, the model has performed well in other NLU tasks such as textual entailment, question paraphrasing, multiple-choice question answering. Finally, to significantly improve summarization, human evaluation was carried out, and student t-test was conducted on the results.

- **ParsBERT** (Farahani et al. 2021) provides monolingual BERT for Persian language, which outperforms multilingual models. Since the Persian language datasets for natural language processing tasks are very small, different datasets have been used for different natural language processing tasks to train the model. The presented model includes five processes; namely data collection, data pre-processing, precise sentence segmentation, pre-training settings, and fine-tuning. The first three steps refer to the data set and the next two refer to the creation of the model. Different data sets (Persian Wikipedia, DigiKala, BingBangPage, etc.) have been used to train the model. Sentiment Analysis and Text Classification are related to Sequence Classification. In short, Sentiment Analysis is known as a special task for Sequence Classification in displaying text emotions. ParsBERT has been evaluated in three downstream tasks: Sentiment Analysis, Text Classification, NER, each of which needs its special dataset. Therefore, several datasets have been used to train this model. The standardization and pre-processing used in the methodology of the article overcome the lack of correct sentences in the Persian language and its complexities. The range of topics and writing styles included in the dataset for pre-training is much more diverse than the multilingual BERT, which only uses the Wikipedia dataset. On the other hand, the multilingual model uses only 70 thousand tokens for 100 languages, while this model has 14 GB of data with more than 3.9 million documents and 100 thousand words. It is lighter and newer than the multilingual model and has better results in downstream tasks such as Sentiment Analysis, Text Classification, and NER.
- **BERT-PersNER** (Jalali Farahani and Ghassem-Sani 2021) presented a language model based on BERT for NRE. It has used transfer learning (Pan and Yang 2010) and active learning (Settles 2009) approaches. A more effective model has been created with the transfer learning approach using limited data that takes less time to train. Moreover, the active learning approach allows for providing a performance very close to the supervised learning method. Also, a conditional random field is used to decode tags in this architecture Instead of only looking for the best tag for each word, it jointly uses neighboring tags to determine a sequence of output tags. It has been evaluated on two datasets; Arman (Poostchi et al. 2018) and Peyma (Shahshahani et al. 2019) with a supervised learning approach.
- **FaBERT** (Masumi et al. 2024) is a new Persian language model based on BERT architecture, trained on the HmBlogs corpus that contains both formal and informal Persian texts. This model has been tested on various NLP tasks including sentiment analysis, named entity recognition, and question answering, showing improved performance while maintaining a compact size. A key advantage of FaBERT is its ability to handle diverse Persian language styles.

Table 4 shows available QA datasets and their evaluation by Language Models briefly. There is a summary of the linguistic models used for each data set in the last column of this table, along with the results.

4.2.2 LLM developments

- Commercial LLMs such as GPT-4 (Achiam et al. 2023), Claude(Caruccio et al. 2024) while primarily trained on English and other high-resource languages, show surprising effectiveness in handling Persian queries and tasks, despite Persian not being a primary

Table 4 QA datasets and language models used

Dataset	Number of questions-answers	Approach	Language models for evaluation		
			LM[a]	EM[b]	F1[c]
(Abadani et al. 2021)	70 K (automatic)	Translation	*ParsBERT*	62.42	65.26
			mBERT	67.73	70.84
			ALBERT	64.71	67.59
(Abadani et al. 2021)	25k (manual)	Native	*ParsBERT*	46.32	50.06
			mBERT	52.86	56.66
			ALBERT	48.11	51.66
(Darvishi et al. 2023)	80k	Native	*ParsBERT*	68.1	82.0
			XLM-RoBERTa	74.8	87.6
			Human	80.3	88.3
(Etezadi and Shamsfard 2020b)	10k	Native	–		
(Jamali et al. 2021)	989 questions 21,915 answers	Community source	*ParsBERT*	–	58.07
			XLM-RoBERTa	–	54.13
			m-BERT	–	50.71
(Kazemi et al. 2022)	20k	Native	*ParsBERT*	73.8	79.08
			ALBERTA-FA	74.9	79.25
			m-BERT	78.8	82.97
			Human	95	96.49
(Babaei Giglou et al. 2022)	16,772 questions	Automated generation	–		
(Kazemi et al. 2023)	500	Native	*ParsBERT*	–	75.61
			BERT	–	74.34
			AlBERT	–	70.93
(Ghafouri et al.,2023)	12,282	Native	*XLM-RoBERTa*	67.33	80.44
			ParsBERT	63.86	75.84
			m-BERT	57.86	72.37
			mT5	21.61	61.28
(Shahriar et al. 2023)	7,704	Native	–		
(ForutanRad et al. 2024)	Small version: 10k questions Large Version: 145k questions	Native	–		

[a]Language mode

[b]Exact match

[c]F-one measure

focus of their training data. However, as a low-resource language, Persian still faces significant challenges in LLM development and application, particularly in tasks requiring deep cultural and linguistic understandings such as poetic text analysis, dialectal variations, complex word formations and cultural references(Abaskohi et al. 2023).

- Recent developments in Persian LLM evaluation have revealed both significant progress and persistent challenges. While models like GPT-4 demonstrate strong capabilities in reasoning and general knowledge tasks, specialized fine-tuned models still outperform them in specific Persian tasks, with some LLMs showing better results when processing translated English rather than direct Persian content(Abaskohi et al. 2023). PersianMMLU (Ghahroodi et al. 2024)s, the Khayyam Challenge, was introduced as a comprehensive framework, offering 20,192 questions across diverse tasks with rich metadata and culturally-aware content. This framework's emphasis on original Persia content rather than translated materials represents an important step toward better understanding and improving LLMs' Persian language capabilities.
- Throughout this study, we will focus on open-source Persian LLMs. Recent significant developments in this area include:
- **PersianMind**(Rostami et al. 2023) builds on LLaMa2-7B-chat (Touvron et al. 2023) by adding 10,000 Persian subwords and training on 2 billion Persian tokens. Using LoRA technique (Hu et al. 2021) for cost-effective training and English Persian parallel datasets to maintain bilingual capabilities, the model achieves GPT-3.5-turbo (OpenAI 2023) comparable performance in reading comprehension and QA tasks. It outperforms mT5-large (Kale 2020) in translation with significantly less training data and generates effective sentence embeddings in both languages, demonstrating superior performance over previous masked language models.
- **PersianLLaMA**(Abbasi et al. 2023) is a Persian language model available in 7B and 13B parameter versions, trained on formal and colloquial Persian texts. The model was evaluated on both generation and understanding tasks using standard benchmarks and automated metrics. it employs two distinct training approaches: one training from scratch using 23 million Persian texts from the OSCAR dataset (Javier et al. 2020), and another using LoRA fine-tuning (Hu et al. 2021) on 2.4 million Persian Wikipedia articles. These different approaches represent varying strategies for adapting LLMs to Persian, with each method utilizing different data sources and training techniques. for evaluation they have incorporated ChatGPT (Using standardized prompts) as an assessment tool for generated Persian texts. The study includes comparisons with published models like mGPT (Shavrina 2022), mT5-XL (Kale 2020),as well as community-developed models like GPT2-Persian[2], and Parsgpt[3]).

5 QA efforts and systems in Persian

Due to the lack of Persian language processing tools, the lack of standard data, and the lack of test data to evaluate the QAS in Persian, few QA systems have been created in Persian language. Due to this, machine learning is a challenging approach to use in these types of

[2] bolbolzaban/gpt2-persian · Hugging Face.

[3] https://huggingface.co/HooshvareLab/gpt2-fa.

systems. Among the studies done on QAS in Persian language, two perspectives can be distinguished. Developing modules of QAS aimed at improvement of these systems and developing QAS as a whole. An overview of these systems is given in Table 5 and the following section summarizes the research in this regard.

In a study (Veisi and Shandi 2020), a dataset on drugs and diseases has been collected. The proposed system has three main modules; query processing, document retrieval, and answer extraction. A sequential architecture has been used for the question processing module, which retrieves the main concept of the question using different components. Rule-based methods, natural language processing, and dictionary-based techniques are used in these components. In the document retrieval module, documents are indexed and searched using the Lucene library. Retrieved documents are ranked using similarity detection algorithms, and the document with the highest rank is selected for the answer extraction module. This module is responsible for finding the most relevant part of the text as an answer. In this research, customized language processing tools such as POS tagger and lemmatizer were also designed for Persian language. The accuracy of this system for 500 sample questions is 83.6%.

Another related study (Tohidi et al. 2021) has considered QA systems with the view that a quality search engine can be considered a semi-mechanized QA system. This approach was created to optimize the performance of web-based QA systems for the Persian language. When the features of the text are extracted, each of them examines the text from different points of view, therefore, a method must be used that considers all these points of view. The proposed approach uses NSGA-II as a Multi-Objective Evolutionary algorithm. This approach has four stages and examines the text from three dimensions:

- Pre-processing the question,
- Retrieving resources, which extracts the set of texts that includes possible answers (in the case with web-based QA systems, it is possible to use web search to extract expressions instead of processing),
- Extracting semantic, lexical, and syntactic features.
- Ranking, where answers with higher probability are extracted using NSGA-II.
- In this paper, the Rasekhoon data set has been used, which contains religious questions and answers.

In another study (Etezadi and Shamsfard 2020a), a knowledge-based approach has been employed to answer complex questions using Farsbase (Sajadi et al. 2018). A system has been proposed based on semantic analysis to convert complex questions to SPARQL. There are three stages in this approach:

1. Finding the entity and limitations from the knowledge graph.
2. Creating a logical form.
3. Ranking and selection.

In this study, Farsbase has been used to answer questions, and PeCoQ (Etezadi and Shamsfard 2020b) has been used to train deep models.

Comparatively to other modules of QAS, the question processing module has been the subject of more studies in Persian. In a study (Razzaghnoori et al. 2018) emphasizes the

Table 5 Studies related to persian question and answering systems based on the researcher's perspective (developing QAS modules or developing of the QAS as a whole)

Research	Perspective	Methods	Architecture	Train and Evaluation dataset
Question classification in Persian using word vectors and frequencies (Razzaghnoori et al. 2018)	Focus on question processing module	Machine learning (SVM, MLP)	-Extracting features -Classification	- University of Tehran Question Dataset 2016 (UTQD.2016) -Translated dataset from an English question dataset
A Knowledge-based Approach for Answering Complex Questions in Persian (Etezadi and Shamsfard 2020a)	Focus on Answer generating Module	Knowledge-based method (graph)	-Finding entity from graph -Creating logical form -Ranking and selection	-Farsbase (Sajadi et al. 2018) and -PeCoQ (Etezadi and Shamsfard 2020b)
Persian question classification using headword and semantic features (Roustaei and Rastegari 2018)	Focus on question processing module	Machine learning (Naïve Bays, KNN, Libsvm, Decision tree)	-Query processing -Feature extraction -Classification	-Manually labeled dataset (queries set of the Hamshahri newspaper and Frequently Asked Questions (FAQs))
Optimizing the Performance of Persian Multi-Objective (Tohidi et al. 2021)	Focus on retrieval module	NSGA-II	-Preprocessing question -Retrieving resources -Extracting features -Ranking	-Rasekhoon question answering dataset (http://www.rasekhoon.net/)
A Persian Medical Question Answering System (Veisi and Shandi 2020)	Focus on developing QAS	–	-Question processing -Retrieval document -Extracting answer	-Manually labeled dataset (medical documents related to drugs and diseases)
FarsNewsQA: A Deep Learning-based Question Answering System for Persian News Articles (Kazemi et al. 2023)	Focus on developing QAS	Deep Learning (BERT-based models)	-Question processing -Answer span prediction -Start/end token estimation	-Training: PersianQuAD (20,000 questions -Evaluation: FarsQuAD (news domain dataset)
PerAnSel: (Mozafari et al. 2022)	Focus on answer selection module	Parallel sequential and transformer-based methods	-Sequential processing -Transformer processing -Answer selection	-PASD (new dataset) -PerCQA -WikiFA

classification of Persian questions. It classifies questions with machine learning approaches. Three methods have been used to extract features. The first method uses clustering algorithms to divide words into clusters and obtain the feature vector related to each question using clustering information. The second method considers each feature vector as a linear combination of query word vectors. It uses the Word2vec method to extract the vector of words. After that, using the TF-IDF method, the previously mentioned linear combination coefficients are adjusted. Then it uses MLP and SVM for classification.

Another study (Roustaei and Rastegari 2018), two features of question's headword and related semantic words have been introduced for the classification of Persian questions, given that if the headword is correctly identified, the accuracy of the answer classification will increase. In general, there are two types of question classification which are different in purpose; Question target classification (QTC) and Subjective Question Classification (SQC). In QTC, the questions are categorized according to the answer categories. It plays a more important role in determining the type of answer, which can be more useful for filtering irrelevant selected answers.

However, in SQC, question classification is carried out according to related subjects, which helps to identify the purpose of question classification. The classification proposed in this paper is based on QTC. Hierarchical classification is used because if the number of classes is considered to be small, the exact classification of questions will not be determined and if the number of classes is large, the calculations will be heavy. First-level classes are few. Each first-level class contains several subclasses. This is a more comprehensive classification. The classes of the first level are called coarse-grained, and the subclasses of the second level are called fine-grained (X. Li and Roth 2002). There are two main approaches in this regard, rule-based and machine learning. Here, machine learning has been employed. Its architecture includes three parts: query preprocessing, feature extraction, and classification using machine learning algorithms.

Another study (Golzari et al. 2022) have proposed an enhanced approach to question classification by combining Gravitational Search Algorithm (GSA) with ensemble classification. Their work focuses on improving question type identification, which they identify as crucial for overall QA system performance.

FarsNewsQA (Kazemi et al. 2023) presents a QAS for Persian news articles, built upon a newly created dataset called FarsQuAD. Analysis revealed that 'What' and 'Who' questions are most frequent in Persian news queries, while 'Why' and 'Which' questions are least common. The system performs best with 'Where' questions and faces challenges with 'What' questions, despite the latter being more frequent in the dataset.

PerAnSel (Mozafari et al. 2022) introduces an innovative approach to Persian answer selection, specifically addressing Persian language's unique challenges (free word order, right-to-left script, morphological richness, and low-resource status). The system combines sequential and transformer-based methods to handle Persian's flexible word ordering. Along with the system, the authors present PASD, large-scale native answer selection dataset for Persian.

6 Discussion and future trends

In close future, search engines will be more based on QAS since users can have access to their natural language queries faster and easier. This can serve as a motivation for researchers to develop these systems in different domains and types. Despite various studies conducted for reviewing research in QAS, most of them are dedicated to the QAS developed in English. There are limited surveys on QAS in other languages, however, still no specific study on Persian QAS research. Although previous efforts

and studies in Persia QAS research are limited, compared to English, research in this area is expanding since the last decade.

In this research, previous studies in Persian QAS are reviewed in terms of datasets, methods, and language models, with particular attention to recent developments in LLMs. While traditional challenges persist, new opportunities and challenges have emerged with the advent of LLMs.

It is challenging to develop question-answering (QA) systems for languages with limited resources, like Persian, for a variety of reasons. Due to a lack of training data for low-resourced languages, the machine learning models are not strong and precise enough. Furthermore, the quality of training data is an essential factor in developing QAS. Deficiency in linguistic resources (annotated text corpora, lexical resources, and anthologies) in Persian, which are applied in natural language tasks such as QAS, results in difficulties in developing models to generate and understand Persian. On the other hand, the complexity of Persian structure and grammar increases ambiguity in texts, making QAS development more difficult.

The implications of our study extend beyond the specific context of Persian QAS and can serve as a valuable guideline for other low-resource languages undertaking the study of question answering systems. While every language presents unique challenges related to its structure and grammar, there are certain challenges that are more general in nature. For example, languages like German or Dutch utilize compound words, and the absence or presence of spaces between the constituent words can alter the overall meaning or languages like Turkish or Finnish have word agglutination where word boundaries play a crucial role in determining meaning. One such challenge is the lack of specialized tools and datasets for these languages. Our research addresses this challenge by highlighting the importance of developing comprehensive language processing tools and creating benchmark datasets specifically tailored to low-resource languages. By emphasizing the significance of resource and tool adaptation, our study encourages collaboration and knowledge sharing among researchers working on question answering systems in low-resource language settings. This collaborative approach can help overcome the common barriers faced by these languages and accelerate progress in the development of effective question answering systems.

According to the previous studies, the following issues need to be addressed for future research:

1. **Traditional Challenges**.

- Lack of benchmark datasets.
- Limited NLP tools compared to CoreNLP(Manning et al. 2014)s.
- Need for larger datasets.

2. **New Challenges in the LLM Era**.

- Need for effective prompt engineering strategies for Persian.
- Evaluation frameworks specific to LLM-based Persian QA.

- Methods to reduce hallucination in Persian responses.
- Integration of cultural context in LLM responses.

3. **Future Directions**:

- Development of Persian-specific LLM evaluation metrics.
- Creation of high-quality instruction tuning datasets.
- Enhanced RAG approaches for Persian knowledge integration.
- Better methods for leveraging multilingual LLM capabilities.

For languages with rich resources, deep learning and LLM techniques are widely used, while traditional machine learning-based techniques remain dominant for languages with limited resources. Although LLMs show promise for Persian QA, significant work remains in creating specialized datasets, evaluation frameworks, and fine-tuning approaches to fully leverage these technologies.

Author contributions Safoura Aghadavoud Jolfaei wrote the main manuscript textAzadeh Mohebi and Safoura Aghadavoud Jolfaei edited and reviewed the text.

Declarations

Competing interests The authors declare no competing interests.

References

Abadani N, Mozafari J, Fatemi A, Nematbakhsh MA, Kazemi A (2021) ParSQuAD: machine translated SQuAD dataset for persian question answering. 2021 7th Int Conf Web Res ICWR 2021 163–168. https://doi.org/10.1109/ICWR51868.2021.9443126

Abaskohi A, Baruni S, Masoudi M, Abbasi N, Babalou MH, Edalat A, Kamahi S, Sani SM, Naghavian N, Namazifard D (2023) Benchmarking large Language models for Persian. A Preliminary Study Focusing on ChatGPT

Abbasi MA, Ghafouri A, Firouzmandi M, Naderi H, Bidgoli BM (n.d.) (eds) PersianLLaMA: Towards Building First Persian Large Language Model. 1–23

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R (2023) "Gpt-4 technical report." arXiv preprint http://arxiv.org/abs/arXiv:2303.08774

Agarwal R (n.d.) (ed) Zero-shot Factual Consistency Evaluation Across Domains. 1–21

Alam F, Chowdhury SA, Boughorbel S, Hasanain M (2024) LLMs for low resource languages in multilingual, multimodal and dialectal settings. EACL 2024–18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of Tutorial Abstracts, 27–33

Alwaneen TH, Azmi AM, Aboalsamh HA, Cambria E, Hussain A (2022) Arabic question answering system: a survey. Artif Intell Rev 55(1):207–253. https://doi.org/10.1007/s10462-021-10031-1

Amirkhani H, AzariJafari M, Faridan-Jahromi S, Kouhkan Z, Pourjafari Z, Amirak A (2023) FarsTail: a Persian natural language inference dataset. Soft Comput. https://doi.org/10.1007/s00500-023-08959-3.
Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, Shakeri S, Taropa E, Bailey P, Chen Z, Chu E, Clark JH, Shafey L, El, Huang Y, Meier-Hellstern K, Mishra G, Moreira E, Omernick M, Robinson K, Wu Y (2023) PaLM 2 Technical Report. May. http://arxiv.org/abs/2305.10403
Antoniou C, Bassiliades N (2022) A survey on semantic question answering systems. 1–42. https://doi.org/10.1017/S0269888921000138
Artetxe M, Ruder S, Yogatama D (2019) On the Cross-lingual transferability of monolingual representations. CoRR, abs/1910.1. http://arxiv.org/abs/1910.11856
Artetxe M, Labaka G, Agirre E (2020) Translation artifacts in cross-lingual transfer learning. In: Proc 2020 Conf Empir Methods Nat Lang Process (EMNLP) 7674–7684. https://doi.org/10.18653/v1/2020.emnlp-main.618
Babaei Giglou H, Beyranvand N, Moradi R, Salehoof AM, Bibak S (2022) ParsSimpleQA: the Persian simple question answering dataset and system over knowledge graph. Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, 59–68. https://aclanthology.org/2022.nlp4dh-1.9
Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72
Berant J, Chou A, Frostig R, Liang P (2013) Semantic Parsing on {F}reebase from Question-Answer Pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1533–1544. https://aclanthology.org/D13-1160
Bordes A, Usunier N, Chopra S, Weston J (2015) Large-scale simple question answering with memory networks. CoRR, abs/1506.0. http://arxiv.org/abs/1506.02075
Bouziane A, Bouchiha D, Doumi N, Malki M (2015) Question answering systems: survey and trends. Procedia - Procedia Comput Sci 73(Awict):366–375. https://doi.org/10.1016/j.procs.2015.12.005
Bui T, Tran O, Nguyen P, Ho B, Nguyen L, Bui T, Quan T (2024) Cross-data knowledge graph construction for LLM-enabled educational question-answering system: a case study at HCMUT. In: Proc 1st ACM Workshop AI-Powered Q&A Syst Multimedia 36–43. https://doi.org/10.1145/3643479.3662055
Calijorne Soares MA, Parreiras FS (2020) A literature review on question answering techniques, paradigms and systems. J King Saud Univ - Comput Inform Sci 32(6):635–646. https://doi.org/10.1016/j.jksuci.2018.08.005
Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G (2024) Intelligent Systems with Applications Claude 2.0 large language model: tackling a real-world classification problem with a new iterative prompt engineering approach. Intell Syst Appl 21(January):200336. https://doi.org/10.1016/j.iswa.2024.200336
Chali Y, Hasan SA, Mojahid M (2015) A reinforcement learning formulation to the complex question answering problem. Inf Process Manag 51(3):252–272. https://doi.org/10.1016/j.ipm.2015.01.002
Chan Y, Oct CL, Pu G, Jenks P (2024) Balancing cost and effectiveness of Synthetic Data Generation Strategies for LLMs. NeurIPS
Chen A, Stanovsky G, Singh S, Gardner M (2019) Evaluating question answering evaluation. MRQA@EMNLP 2019 - Proceedings of the 2nd Workshop on Machine Reading for Question Answering, 119–124. https://doi.org/10.18653/v1/d19-5817
Cimiano P, Lopez V, Unger C, Cabrio E, Ngonga Ngomo A-C, Walter S (2013) Multilingual question answering over Linked Data (QALD-3): lab overview. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information Access evaluation. Multilinguality, multimodality, and visualization. Springer, Berlin, pp 321–332
Clark E, Celikyilmaz A, Smith NA (2019) Sentence Mover{'}s similarity: automatic evaluation for multi-sentence texts. In: Proc 57th Annual Meeting Association Comput Linguistics 2748–2760. https://doi.org/10.18653/v1/P19-1264
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. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747
Darvishi K, Shahbodaghkhan N, Abbasiantaeb Z, Momtazi S (2023) PQuAD: A Persian question answering dataset. Comput Speech Lang 80:101486. https://doi.org/10.1016/j.csl.2023.101486
Daull, X., Bellot, P., Bruno, E., Martin, V., & Murisasco, E. (2023). Complex QA and language models hybrid architectures, Survey. arXiv preprint http://arxiv.org/abs/arXiv:2302.09051.
Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019–2019 Conf North Am Chapter Association Comput Linguistics: Hum Lang Technol - Proc Conf 1Mlm:4171–4186

Dimitrakis E, Sgontzos K, Tzitzikas Y (2019) A survey on question answering systems over linked data and documents
Dimitrakis E, Sgontzos K, Tzitzikas Y (2020) A survey on question answering systems over linked data and documents. J Intell Inform Syst 55(2):233–259. https://doi.org/10.1007/s10844-019-00584-7
Do H, Ostrand R, Weisz J, Dugan C, Sattigeri P, Wei D, Murugesan K, Geyer W (2024) Facilitating human-LLM collaboration through factuality scores and source attributions. https://doi.org/10.48550/arXiv.2405.20434
Elfadil SSAN, Jarajreh M, Algarni S (2021) Question Answering systems: a systematic literature review. Int J Adv Comput Sci Appl 12(3):495–502. https://doi.org/10.14569/IJACSA.2021.0120359
Etezadi R, Shamsfard M (2020a) A Knowledge-based approach for answering complex questions in Persian
Etezadi R, Shamsfard M (2020b) PeCoQ: A dataset for persian complex question answering over knowledge graph. In: 2020 11th International Conference on Information and Knowledge Technology, IKT 2020, 102–106. https://doi.org/10.1109/IKT51791.2020.9345610
Etezadi R, Karrabi M, Zare N, Sajadi MB, Pilehvar MT (2022) DadmaTools: a natural language processing toolkit for the Persian language. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, 124–130
Farahani M, Gharachorloo M, Farahani M, Manthouri M (2021) ParsBERT: transformer-based model for Persian language. Neural Process Lett 53(6):3831–3847. https://doi.org/10.1007/s11063-021-10528-4
Farea A, Yang Z, Duong K, Perera N, Emmert-Streib F (2022) Evaluation of question answering systems: complexity of judging a natural language. http://arxiv.org/abs/2209.12617
Fayyaz M, Yin F, Sun J, Peng N, Angeles L (2024) Evaluating human alignment and model faithfulness of LLM rationale
Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur A, Lally A, Murdock JW, Nyberg E, Prager J, Schlaefer N, Welty C (2010) Building Watson: an overview of the DeepQA project. AI Magazine 31:59–79
ForutanRad J, HourAli M, KeyvanRad M (2024) Farsi question and answer dataset (FarsiQuAD). Signal Data Process 20(4). https://doi.org/10.61186/jsdp.20.4.107.
Franco W, Avila CVAOS, Maia G, Brayner A, Vidal VMP, Carvalho F, Pequen V (2020) Ontology-based question answering systems over knowledge bases: a survey. In: 22nd International Conference on Enterprise Information Systems (ICEIS 2020), 532–539
Fu Y, Ou L, Chen M, Wan Y, Peng H, Khot T (2023) Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance. arXiv preprint http://arxiv.org/abs/arXiv:2305.17306.
Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H (2023) Retrieval-augmented generation for large language models: A survey. arXiv preprint http://arxiv.org/abs/2312.10997
Ghafouri A, Naderi H, Firouzmandi M (2023) IslamicPCQA: A Dataset for Persian Multi-hop Complex Question Answering in Islamic Text Resources. arXiv preprint http://arxiv.org/abs/2304.11664
Ghahroodi O, Nouri M, Sanian MV, Sahebi A, Dastgheib D (2024) Khayyam challenge (PersianMMLU). Is Your LLM Truly Wise
Golzari S, Sanei F, Saybani MR, Harifi A, Basir M (2022) Question classification in question answering system using combination of ensemble classification and feature selection. J AI Data Min 10(1):15–24. https://doi.org/10.22044/JADM.2021.10016.2142
GPT-4 Technical Report (2023) 4, 1–100
Habib MK (2021) The challenges of Persian user-generated textual content: a machine learning-based approach. 1–12. http://arxiv.org/abs/2101.08087
Heie MH, Whittaker EWD, Furui S (2012) Question answering using statistical language modelling. Comput Speech Lang 26(3):193–209. https://doi.org/10.1016/j.csl.2011.11.001
Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Chen W (2021) LoRA: low-rank adaptation of large language models. CoRR, abs/2106.09685. https://arxiv.org/abs/2106.09685
Huang Q, Bu J, Xie W, Yang S, Wu W, Liu L (2019) Multi-task sentence encoding model for semantic retrieval in question answering systems. July, pp 1–8
Jalali Farahani F, Ghassem-Sani G (2021) BERT-PersNER: a new model for Persian named entity recognition. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, 647–654
Jamali N, Yaghoobzadeh Y, Faili H (2021) PerCQA: Persian community question answering dataset. http://arxiv.org/abs/2112.13238
Javier P, Suárez O, Romary L, Sagot B (2020) A monolingual approach to contextualized word embeddings for mid-resource languages. 1703–1714
Jurafsky D, Martin JH (2020) Speech and language processing
Kale M (2020) mT5: a massively multilingual pre-trained text-to-text transformer

Kazemi A, Mozafari J, Nematbakhsh MA (2022) PersianQuAD: the native question answering dataset for the Persian language. IEEE Access 10:26045–26057. https://doi.org/10.1109/ACCESS.2022.3157289
Kazemi A, Zojaji Z, Malverdi M, Mozafari J, Ebrahimi F, Abadani N, Varasteh MR, Nematbakhsh MA (2023) FarsNewsQA: a deep learning-based question answering system for the persian news articles. Inform Retr J 26(1). https://doi.org/10.1007/s10791-023-09417-2
Khandelwal A, Singh H, Gu H, Chen T, Zhou K (2024) A simple contrastive learning based approach
Khvalchik M, Galkin M (2020) El Departamento de Nosotros: how machine translated corpora affects language models in MRC tasks. ArXiv. https://doi.org/10.48550/arxiv.2007.01955
Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) ALBERT: a lite BERT for self-supervised learning of language representations. ArXiv. https://doi.org/10.48550/arXiv.1909.11942
Lan Y, Li X, Liu X, Li Y, Qin W, Qian W (2023) Improving zero-shot visual question answering via large language models with reasoning question prompts. MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia, 4389–4400. https://doi.org/10.1145/3581783.3612389
Latifi M (2018) Using natural language processing for question answering in closed and open domains. TDX (Tesis Doctorals En Xarxa), March, 1–165
Leite B, Cardoso HL (2024) In Narrative comprehension a.
Lewis P, Oguz B, Rinott R, Riedel S, Schwenk H (2020) MLQA: evaluating cross-lingual extractive question answering. 7315–7330. https://doi.org/10.18653/v1/2020.acl-main.653
Li X, Roth D (2002) Learning question classifiers. {COLING} 2002: The 19th International Conference on Computational Linguistics. https://aclanthology.org/C02-1150
Li T, Ma X, Zhuang A, Gu Y, Su Y, Chen W (2023) Few-shot In-context learning for knowledge base question answering. 1, 6966–6980
Li Z, Fan S, Gu Y, Li X, Duan Z, Dong B, Liu N, Wang J (2024) FlexKBQA: a flexible LLM-powered framework for few-shot knowledge base question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18608–18616. https://doi.org/10.1609/aaai.v38i17.29823
Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out Post-Conference Workshop of ACL, pp. 74-81 https://aclanthology.org/W04-1013.pdf
Liu T, Zhang Y, Brockett C, Mao Y, Sui Z, Chen W, Dolan B (2019) A token-level Reference-. free Hallucination Detection Benchmark for Free-form Text Generation
Liu X, Lai H, Yu H, Xu Y, Zeng A, Du Z, Zhang P, Dong Y, Tang J (2023) WebGLM: towards an efficient web-enhanced question answering system with human preferences. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 4549–4560. https://doi.org/10.1145/3580305.3599931
Lopez V, Unger C, Cimiano P, Motta E (2013) Evaluating question answering over linked data. J Web Semant 21:3–13. https://doi.org/10.1016/j.websem.2013.05.006
Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. https://doi.org/10.3115/v1/P14-5010
Masumi M, Majd SS, Shamsfard M, Beigy H, Science C (2024) FaBERT: Pre-training BERT on persian Blogs
Mishra A, Jain SK (2016) A survey on question answering systems with classification. J King Saud Univ - Comput Inform Sci 28(3):345–361. https://doi.org/10.1016/j.jksuci.2014.10.007
Mishra AK, Saxena PA (2019) Survey on question and answering retrieval system using Hadoop. 1100–1103
Mozafari J, Kazemi A, Moradi P, Nematbakhsh MA (2022) PerAnSel: a novel deep neural network-based system for Persian question answering. Comput Intell Neurosci 2022(1). https://doi.org/10.1155/2022/3661286
Neves M, Leser U (2015) Question answering for biology. Methods 74:36–46. https://doi.org/10.1016/j.ymeth.2014.10.023
Ojokoh B, Adebisi E (2019) Rev Question Answering Syst 17:717–758. https://doi.org/10.13052/jwe1540-9589.1785
OpenAI (2023) GPT turbo 3.5. https://platform.openai.com/docs/models/gpt-3-5
Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
Papineni K, Roukos S, Ward T, Zhu W-J (2002) {B}leu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135
Poostchi H, Borzeshi Z, E., Piccardi M (2018) BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} 2018)
Prabhu VVD, Anand A (2024) DEXTER: a Benchmark for open-domain Complex question answering using LLMs. 1–22. http://arxiv.org/abs/2406.17158

Pundge AM (2016) Question answering system, approaches techniques: a review 141(3):34–39
Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) {S}tanza: a Python Natural Language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14
Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuad: 100,000+questions for machine comprehension of text. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2383–2392. https://doi.org/10.18653/v1/d16-1264
Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: unanswerable questions for SQuAD. ArXiv, pp 784–789
Razzaghnoori M, Sajedi H, Jazani IK (2018) Question classification in Persian using word vectors and frequencies. Cogn Syst Res 47:16–27. https://doi.org/10.1016/j.cogsys.2017.07.002
Rogers A, Kovaleva O, Rumshisky A (2020) A primer in bertology: what we know about how bert works. Trans Assoc Comput Linguistics 8:842–866. https://doi.org/10.1162/tacl_a_00349
Rostami P, Salemi A, Dousti MJ (2023) PersianMind: a cross-Lingual Persian-English Large Language Model
Roustaei A, Rastegari H (2018) Persian question classification using headword and semantic features. J Theoretical Appl Inform Technol 96(21):7206–7214
Saini A, Yadav PK (2017) A Survey on question–answering system. Int J Eng Comput Sci 6(3):20453–20457. https://doi.org/10.18535/ijecs/v6i3.09
Sajadi MB, Bidgoli BM, Hadian A (2018) FarsBase: a cross-domain farsi knowledge graph. CEUR Workshop Proceedings, 2198, 4–7
Salemi A, Kebriaei E, Neisi Minaei G, Shakery A (2021) ARMAN: pre-training with semantically selecting and reordering of sentences for Persian Abstractive Summarization. 9391–9407. https://doi.org/10.18653/v1/2021.emnlp-main.741
Saxena Y, Chopra S, Tripathi AM (2024) Evaluating consistency and reasoning capabilities of Large Language Models
Schmidt M, Bartezzaghi A, Vu NT (2024) Prompting-based synthetic data generation for few-shot question answering. 13168–13178
Settles B (2009) Computer aciences active learning literature survey. January
Shahriar et al (2023) MeDiaPQA: a question-answering dataset on Persian medical dialogues. Mendeley Data, V1. https://doi.org/10.17632/k7tzmrhr6n.1
Shahshahani MS, Mohseni M, Shakery A, Faili H (2019) PAYMA: a tagged corpus of Persian named entities. Signal Data Process 16(1). https://doi.org/10.29252/jsdp.16.1.91
Shamsfard M (2019) Challenges and Opportunities in Processing Low Resource Languages: a study on Persian. International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide, 291–295. https://lt4all.org/media/papers/O9/168.pdf
Shavrina T (2022) mGPT: Few-shot learners go Multilingual. ii
Shuster K, Poff S, Chen M, Kiela D, Weston J (2021) Retrieval augmentation reduces hallucination in conversation. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3784–3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320
Snell R (2009) The Indo-Aryan languages. By Colin P. Masica. (Cambridge Language Surveys.) pp. xvi, 539. Cambridge etc., Cambridge University Press, 1991. £65.00. Journal of the Royal Asiatic Society, 4, 284–285. https://doi.org/10.1017/S1356186300005678
Tam D (2023) Evaluating the factual consistency of large language models through news summarization. 5220–5255
Thakur N, Reimers N, Ruckl'e A, Srivastava A, Gurevych I (2021) BEIR: a heterogenous benchmark for zero-shot evaluation of Information Retrieval models. ArXiv, abs/2104.0. https://api.semanticscholar.org/CorpusID:233296016
Tohidi N, Dadkhah C, Rustamov RB (2021) Optimizing persian multi-objective question answering system. Int J Tech Phys Probl Eng 13(46):62–69
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y.,... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst 2017-Decem(Nips), 5999–6009
Veisi H, Shandi HF (2020) A Persian Medical question answering system. 29(6):1–29. https://doi.org/10.1142/S0218213020500190
Wang C, Xu Y, Peng Z, Zhang C, Chen B, Wang X, Feng L, An B (2023a) Keqing: knowledge-based question answering is a nature chain-of-thought mentor of LLM. http://arxiv.org/abs/2401.00426

Wang Y, Ma X, Chen W (2023b) Augmenting black-box LLMs with medical textbooks for clinical question answering. http://arxiv.org/abs/2309.02233
Wang J, Yang Z, Yao Z, Yu H (2024) JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability. http://arxiv.org/abs/2402.17887
Wei J, Schuurmans D, Chi EH, Le QV, Jan CL (2022) Chain-of-thought prompting elicits reasoning in large Language models. NeurIPS, pp 1–43
Wolfson T, Bogin B, Katz U, Deutch D, Berant J (2023) Answering questions by meta-reasoning over multiple chains of thought. 5942–5966
Yang Z, Qi P, Zhang S, Bengio Y, Cohen W, Salakhutdinov R, Manning CD (2018) {H}otpot{QA}: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369–2380. https://doi.org/10.18653/v1/D18-1259
Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2016). A survey of intelligent question answering system using nlp and information retrieval techniques. International Journal of Advanced Research in Computer and Communication Engineering, 5(5), 536-540. https://doi.org/10.17148/IJARCCE.2016.55134
Zaib M, Zhang WE, Sheng QZ, Mahmood A, Zhang Y (2021) Conversational question answering: a survey. http://arxiv.org/abs/2106.00874
Zan D, Chen B, Zhang F, Lu D, Wu B, Guan B, Wang Y, Lou JG (2023) Large Language Models Meet NL2Code: a survey. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1(2), 7443–7464. https://doi.org/10.18653/v1/2023.acl-long.411
Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) BERTScore: evaluating text generation with BERT. ArXiv, abs/1904.0. https://doi.org/10.48550/arXiv.1904.09675
Zhang Y, Chen Z, Fang Y, Lu Y, Li F, Zhang W, Chen H (2023) Knowledgeable preference alignment for LLMs in domain-specific question answering. http://arxiv.org/abs/2311.06503
Zhao, Z., Jin, Q., Chen, F., Peng, T., & Yu, S. (2022). Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. arXiv preprint arXiv:2202.13876. https://arxiv.org/abs/2202.13876
Zhao W, Liu Y, Niu T, Wan Y, Yu PS, Joty S, Zhou Y, Yavuz S (2024) DIVKNOWQA: Assessing the Reasoning Ability of LLMs via Open-Domain Question Answering over Knowledge Base and Text. Findings of the Association for Computational Linguistics: NAACL 2024 - Findings, 51–68. https://doi.org/10.18653/v1/2024.findings-naacl.5