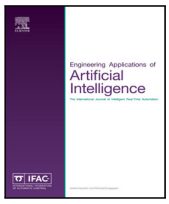
Engineering Applications of Artificial Intelligence 156 (2025) 110997

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survey paper

Understanding question-answering systems: Evolution, applications, trends, and challenges

Amer Farea [a], Frank Emmert-Streib [a,*]

[a] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, 33720, Finland

ARTICLE INFO

Keywords:
Neural networks
Machine learning
Natural language processing
Extractive Question Answering
Generative Question Answering
Visual Question Answering

ABSTRACT

Question answering (QA) systems have garnered significant attention in recent years due to their potential to bridge the gap between human language understanding and machine intelligence. Consequently, a wide variety of approaches have been developed, each tailored to specific tasks. In this survey paper, we provide a comprehensive overview of three prominent QA paradigms: Extractive , generative, and Visual QA. We discuss the underlying principles, methodologies, applications, challenges, and recent trends in each of these areas. By synthesizing insights from the existing literature and research findings, we aim to provide a holistic understanding of extractive, generative, and Visual QA systems and offer insights into their strengths, limitations, and future directions.

1. Introduction

The burgeoning field of question-answering (QA) systems highlights the evolving synergy between human language comprehension and Artificial Intelligence (AI) (Soares and Parreiras, 2020; Jin et al., 2022). These systems, designed to swiftly furnish accurate and relevant responses to queries posed in natural language, epitomize the convergence of advanced computational techniques and the intricacies of human communication. However, they also face challenges, including limitations in understanding context, handling ambiguity, and addressing biases in the data they are trained on, which can impact their accuracy and fairness.

From streamlining information retrieval processes to powering virtual assistants and enhancing educational tools, QA systems permeate myriad domains, catalyzing transformative changes in human–computer interaction (HCI) (Manning, 2008; Voorhees et al., 2005; Cui et al., 2019; Zhang et al., 2021).

At the heart of every QA endeavor lies the fundamental quest to discern the underlying intent and semantics of user queries and, subsequently, to retrieve or generate cogent responses from a plethora of information sources (Rajpurkar et al., 2018; Fader et al., 2014; Weston et al., 2015). This introductory section serves as a gateway to the nuanced world of QA systems, delving into the intricate processes and methodologies that underpin their functionality.

The subsequent overview outlines the key paradigms of QA systems, each offering a unique approach to addressing the complex challenges of natural language understanding (NLU). Extractive Question Answering (EQA), similar to meticulous librarians combing through extensive archives, focuses on identifying and extracting relevant information directly from text sources (Wang et al., 2022; Demner-Fushman and Lin, 2006). By leveraging methods rooted in information retrieval (IR) and natural language processing (NLP), these systems are particularly adept at locating specific passages or sentences that contain answers to user queries, making them invaluable for tasks like reading comprehension and fact verification.

Generative Question Answering (GQA): In contrast, GQA systems embody the art of linguistic creation, dynamically synthesizing responses based on an understanding of query semantics and contextual cues (Lewis and Fan, 2018). By harnessing the power of advanced natural language generation (NLG) models, these systems offer unparalleled flexibility in generating responses to queries, paving the way for sophisticated dialogue systems, creative writing aids, and conversational agents. For example, the recent advancements in ChatGPT-4 (OpenAI, 2023) and Gemini (Google DeepMind, 2024) (Masalkhi et al., 2024) have enabled more human-like and context-aware interactions, expanding their use in virtual healthcare assistants that offer personalized advice and emotional support (e.g., Woebot Health, 2023) (Olawade et al., 2024). Additionally, models like PaLM 2 (Google, 2023) (Chowdhery et al., 2023) have made breakthroughs in multilingual content creation, facilitating real-time translation and content generation across diverse languages, further enhancing the potential applications of NLG systems.

* Corresponding author.
E-mail address: frank.emmert-streib@tuni.fi (F. Emmert-Streib).

https://doi.org/10.1016/j.engappai.2025.110997
Received 15 August 2024; Received in revised form 26 March 2025; Accepted 24 April 2025
Available online 22 May 2025

Visual Question Answering (VQA): The emergence of VQA systems marks a paradigm shift in QA, as these systems adeptly integrate visual information, such as images or videos, with textual queries to provide holistic responses (Antol et al., 2015). Leveraging multimodal fusion techniques, VQA systems unravel the intricacies of visual content, enabling tasks such as image understanding, scene analysis, and content-based image retrieval. These systems hold immense promise in domains ranging from assistive technology for the visually impaired to immersive augmented reality experiences.

Furthermore, this paper categorizes QA systems into distinct types based on their underlying mechanisms and functionalities, elucidating the unique strengths and limitations of each approach (Al Mahmud et al., 2020; Riloff and Thelen, 2000; Manna et al., 2020; Putra et al., 2016; Abdel-Nabi et al., 2023; Sarrouti and El Alaoui, 2017; Park et al., 2021; Cortes et al., 2018; Pakray et al., 2011; Bhaskar et al., 2013). By navigating this taxonomy, our goal is to provide a comprehensive understanding of the varied landscape of QA systems, enabling researchers and practitioners to effectively choose and customize solutions according to specific use cases and challenges.

The overarching goal of this paper is to provide a roadmap for navigating the intricate terrain of QA systems, from foundational principles to cutting-edge advancements. By elucidating the core paradigms, methodologies, and applications of EQA, generative, and VQA systems, this introduction sets the stage for a deeper exploration of the evolving landscape of HCI and NLU.

The subsequent sections of the paper are organized as follows: Section 2 outlines the motivations behind this study, aiming to provide a holistic understanding of three prominent QA paradigms—EQA, GQA, and VQA—while highlighting their strengths, limitations, and future directions. Section 3 offers an overview of the study. Section 4 offers an overview of EQA, Section 5 explores GQA, Section 6 centers on VQA, Section 7 conducts a Comparative Analysis of QA Paradigms, Section 8 addresses Challenges and Future Directions, and lastly, Section 9 presents the Conclusion.

2. Motivation

Studying EQA, GQA, and VGA systems is motivated by the need to address diverse challenges in NLU and HCI.

EQA systems, which retrieve answers directly from text, are essential for efficiently accessing relevant information from large corpora. However, the accuracy of these systems heavily depends on the quality of the data sources. If the underlying corpus contains outdated, biased, or incomplete information, the system's outputs may be inaccurate or misleading. For example, if an EQA system relies on a dataset that contains historical inaccuracies or unverified claims, it could inadvertently retrieve and present incorrect answers, undermining the reliability of the system. Therefore, understanding and improving these systems can enhance IR processes, search engines, and text summarization tools, thereby streamlining access to vast amounts of textual data. GQA systems, on the other hand, offer the capability to generate responses dynamically, allowing for more flexible and nuanced interactions between users and machines. Researching GQA can lead to advancements in dialogue systems, virtual assistants, and NLG, enabling richer and more human-like interactions. Additionally, the emergence of VQA systems stems from the increasing prevalence of multimedia content and the growing importance of visual understanding in AI applications. By studying VQA, researchers aim to develop AI systems capable of interpreting and reasoning about visual information, leading to innovations in fields such as image understanding, augmented reality, and accessibility technologies for the visually impaired. In summary, studying EQA, GQA, and VQA systems is motivated by the broader goal of advancing NLU, NLG, enhancing HCI, and enabling AI systems to effectively process and respond to diverse forms of information.

3. Overview

Question-answering (QA) systems are designed to automatically provide users with accurate, relevant, and concise answers to their questions, typically formulated in natural language. These systems play a crucial role in bridging the gap between human language understanding and machine intelligence, enabling users to interact with computers in a more intuitive and natural manner. QA systems have applications across various domains, including information retrieval, education, customer support, virtual assistants, and more (Manning, 2008; Voorhees et al., 2005; Cui et al., 2019; Ferrucci et al., 2010; Zhang et al., 2021).

The primary objective of QA systems is to comprehend the meaning and intent behind a user's question and retrieve or generate a suitable and coherent response from various sources (e.g., a knowledge base or corpus of information) (Rajpurkar et al., 2018; Fader et al., 2014; Weston et al., 2015). The key components and processes involved in a QA system typically include:

- Natural Language Understanding: The system processes and analyzes the natural language input (questions) to understand the semantics (Li and Yang, 2018), intent (Weld et al., 2022), and context of the queries (Bates, 1995). This may involve techniques such as tokenization (Ly et al., 2020; Fares et al., 2013), part-of-speech tagging (Chiche and Yitagesu, 2022), syntactic parsing (Manning and Schutze, 1999), semantic analysis (Maulud et al., 2021), and named entity recognition (Nadeau and Sekine, 2007).
- Information Retrieval: The system retrieves relevant information from structured or unstructured data sources, such as databases (Ostrovsky and Skeith III, 2007), documents (Mitra and Chaudhuri, 2000), web pages (Kobayashi and Takeda, 2000), or knowledge bases (Diefenbach et al., 2018). This step involves searching and selecting relevant passages, or documents that contain potential answers to the user's question (Manning and Schutze, 1999).
- Answer Extraction or Generation: Depending on the QA architecture, the system either extracts the answer directly from the selected sources or retrieved passages (EQA) or generates a new answer in natural language (GQA) (Wang et al., 2022; Pandya and Bhatt, 2021; Lewis and Fan, 2018). EQA systems typically identify and select the most relevant passages or sentences containing the answer, while GQA systems synthesize new text to form coherent responses.
- Scoring and Ranking: The system evaluates and scores potential answers based on factors such as relevance (Zhong et al., 2022), coherence (Shah et al., 2019b), confidence (Guo et al., 2017; Tian et al., 2023), and correctness (de Winter, 2023). In EQA, this may involve ranking candidate passages or sentences based on their relevance to the question (Pezeshkpour, 2023). In GQA, the system may score generated responses based on their semantic coherence and grammatical correctness (Sun et al., 2023).
- Presentation of Results: The system presents the final answer or a ranked list of candidate answers to the user in a human-readable format. This may include text-based responses, structured data, visualizations, or multimedia content, depending on the nature of the question and the output format supported by the system.

Overall, a QA system aims to provide users with accurate, relevant, and concise answers to their questions, leveraging techniques from NLP, IR, Machine Learning (ML), and other related fields to achieve this goal. QA systems can be implemented using various approaches depending on their functionality and underlying technologies. Here are some common types:

- Keyword-based Question Answering Systems: These systems match keywords or key phrases from the user's question with relevant documents or passages in a corpus to find the answer (Al Mahmud et al., 2020). They are relatively simple and often rely on techniques like IR and keyword matching.
- Rule-based Question Answering Systems: Rule-based systems use predefined rules or patterns to extract information from the input question and find the answer (Riloff and Thelen, 2000). These rules can be based on the syntactic or semantic structures of the language. They are effective for answering specific types of questions but may lack flexibility (Chiticariu et al., 2013). However, pattern matching techniques guided by predefined rules are employed to extract information from a provided dataset or knowledge repository in order to formulate responses to user inquiries. Nonetheless, while effective in certain cases, rule-based pattern matching methods lack the intelligence to consistently grasp the contextual nuances of queries, resulting in static and occasionally inappropriate responses (Chen and Zulkernine, 2021).
- Information Retrieval-based Question Answering Systems: These systems retrieve relevant information from large collections of documents or databases to answer questions (Manna et al., 2020). They typically use techniques like indexing (Tsatsaronis et al., 2015), ranking (Lee et al., 2018), and retrieval to find the most relevant documents or passages containing the answer.
- Semantic-based Question Answering Systems: Semantic-based systems (Putra et al., 2016) focus on understanding the meaning of the question and the content of the documents to provide more accurate answers. They often use NLP techniques like named entity recognition (Nadeau and Sekine, 2007), syntactic and semantic parsing (Manning and Schutze, 1999; Maulud et al., 2021), and entity linking (Shen et al., 2014) to analyze the question and retrieve relevant information.
- Machine Learning-based Question Answering Systems: These systems use ML algorithms (Abdel-Nabi et al., 2023; Sarrouti and El Alaoui, 2017) to learn patterns from large datasets of questions and answers. They can be trained on labeled datasets to predict answers to new questions. Deep learning models such as Recurrent Neural Networks (RNNs) (Serban et al., 2016a), Convolutional Neural Networks (CNNs) (Li et al., 2021), and transformer models (Kumari and Pushphavati, 2022) (like Bidirectional Encoder Representations from Transformers (BERT) (Qu et al., 2019) and Generative Pre-trained Transformer (GPT) (Klein and Nabi, 2019)) are often used in these systems.
- Knowledge Graph-based Question Answering Systems: Knowledge graph-based systems represent knowledge in the form of a graph (Park et al., 2021), where entities are nodes and relationships are edges. These systems use graph-based algorithms to navigate the knowledge graph and find answers to questions by traversing relationships between entities.
- Hybrid Question Answering Systems: Hybrid systems combine multiple approaches mentioned above to leverage the strengths of each (Cortes et al., 2018; Pakray et al., 2011; Bhaskar et al., 2013). For example, a system may use keyword matching for simple questions and switch to ML-based techniques for more complex questions.

However, each pe of QA system has its strengths and limitations, and the choice of system depends on factors such as the complexity of questions, the size of the knowledge base, and the desired accuracy of answers. Depending on the approach and methodology employed, QA systems can be classified into several categories, each with its own strengths, limitations, and areas of application. Some of the key paradigms, as illustrated in Fig. 1, include:

- Extractive Question Answering: EQA systems (Wang et al., 2022), a subset of information retrieval-based systems, are designed to identify and extract relevant information directly from a given set of documents or passages. By leveraging techniques from IR and NLP, these systems (i.e., Index vector (Manning, 2008), BM25 (Robertson et al., 2009), BERT (Devlin et al., 2018)) locate specific passages containing the answer to a given question, making them particularly effective for factoid questions where the answer can be directly inferred from the text (Demner-Fushman and Lin, 2006). Unlike GQA systems, which generate new text, EQA systems operate by extracting answers directly from the context or document. They are commonly used in tasks such as reading comprehension and fall under the broader categories of rule-based or ML-based systems, depending on the methodologies used for answer identification and extraction.
- Generative Question Answering: GQA systems operate by synthesizing new text in response to user queries (Lewis and Fan, 2018), leveraging advanced NLG techniques like sequence-to-sequence modeling (Yin et al., 2015). Unlike EQA systems, which retrieve answers from pre-existing passages, GQA systems (i.e., T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), BART (Lewis, 2019)) dynamically generate responses based on comprehension of the question and context (Liu et al., 2019; Song et al., 2017). These systems are adept at handling tasks requiring open-domain reasoning (Cheng et al., 2021), dialogue generation (Shen et al., 2021; Honovich et al., 2021), and complex language understanding (Liu et al., 2019), where answers may not be readily available in the provided text. GQA falls within the realm of ML-based systems, as it often involves training models to generate answers from input questions and contexts, while also incorporating elements of semantic-based systems (Niu et al., 2021) to understand question semantics and produce coherent responses.
- Visual Question Answering: VQA systems represent an advancement in traditional QA by integrating visual information (Antol et al., 2015), such as images or videos (Wu et al., 2017), with textual queries. These systems (i.e., VQA (Antol et al., 2015), VilBERT (Lu et al., 2019), VisualGPT (Chen et al., 2022)) employ multimodal fusion techniques (Lu et al., 2023) to merge visual and textual inputs, enabling them to generate answers that require an understanding of both modalities. VQA finds applications in various domains, including image understanding (Goyal et al., 2017), scene analysis (Biten et al., 2019; Kafle and Kanan, 2017a), and assistive technology (Das et al., 2019), empowering users to pose questions about visual content. By combining computer vision and NLP techniques (Wu et al., 2017), VQA systems can interpret images and understand questions, enabling tasks such as object recognition (Ramakrishnan et al., 2017), scene understanding, and spatial reasoning (Azuma et al., 2022). These systems have diverse applications, including image captioning (Anderson et al., 2018), assisting visually impaired individuals (Antol et al., 2015), and content-based image retrieval (Yu et al., 2020). VQA is a specialized type of QA system that processes both visual and textual inputs, making it a hybrid approach within the broader category of multimodal question-answering systems (Cadene et al., 2019; Ben-Younes et al., 2017; Lu et al., 2023).

These systems leverage a combination of information retrieval and ML techniques to process visual and textual data and generate accurate responses. For instance, recent innovations such as multimodal transformers—like Contrastive Language-Image Pre-training (CLIP) by OpenAI (Radford et al., 2021) and Flamingo by DeepMind (Alayrac et al., 2022)—combine vision and language processing into a single architecture, enabling more nuanced and context-aware responses. These systems can not only retrieve and understand textual data but also interpret images, videos, and even real-time sensor data. In practical applications, such

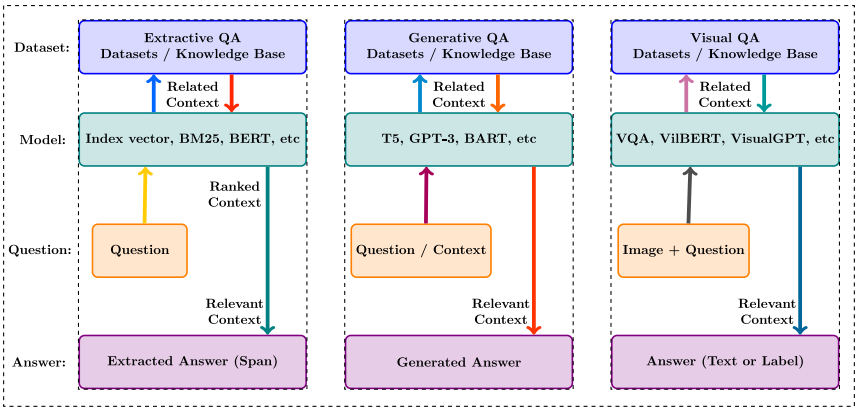


Fig. 1. Overview of three prominent QA paradigms: EQA, GQA, and VQA. EQA retrieves exact spans of text from a given passage, while GQA generates novel answers that may not appear verbatim in the source text. VQA extends traditional QA by incorporating multimodal inputs, combining text-based questions with images, and requiring the model to understand both textual and visual elements for accurate responses. VQA can adopt both EQA and GQA, but the key distinction lies in the use of visual content in addition to text, introducing additional complexity. This classification highlights the differences in input modalities (text-only vs. multimodal) and answering strategies (EQA vs. GQA), with VQA being a unique challenge due to its multimodal nature.

as autonomous vehicles (Talpes et al., 2020; Nordhoff et al., 2023) and augmented reality (AR) platforms (Silva et al., 2023), this combination allows for real-time decision-making based on both visual and textual inputs, greatly improving accuracy and adaptability in dynamic environments.

Overall, the three prominent QA paradigms — EQA, GQA, and VQA — highlight distinct problem-solving strategies, as well as different input and output formats. EQA retrieves exact spans of text from a passage to answer a question, relying on IR from the source material. GQA, on the other hand, generates novel answers that may not appear verbatim in the text, often requiring deeper reasoning. VQA extends traditional QA by incorporating both text-based questions and images, requiring the model to understand both textual and visual content for accurate responses. This framework emphasizes the differences in input modalities (text-only vs. multimodal) and answering strategies (EQA vs. GQA), with VQA introducing additional complexity due to the need for multimodal reasoning. VQA can leverage both EQA and GQA, making it a unique challenge compared to text-only QA systems.

Furthermore, QA systems are evaluated based on various scores, including accuracy (Singhal et al., 2023), precision, recall, fluency (Van Der Lee et al., 2019), and relevance (Zhong et al., 2022). Benchmark datasets, such as SQuAD (Rajpurkar et al., 2016), COCO-QA (Lin et al., 2014), and VQA (Zellers et al., 2018), provide standardized evaluation tasks for assessing the performance of QA systems across different domains and modalities.

Overall, QA systems continue to evolve rapidly, driven by advancements in AI, ML, and NLP. By enabling seamless interaction between humans and machines, QA systems contribute to the development of intelligent applications and services that enhance productivity, accessibility, and user experience across diverse domains.

The primary objective of this survey paper is to provide a comprehensive overview of EQA, GQA, and VQA systems, exploring their methodologies, applications, and advancements. The paper aims to offer insights into the evolution of QA systems, from traditional EQA to more advanced GQA and VQA paradigms, highlighting the strengths, limitations, and challenges associated with each approach.

By synthesizing existing research and literature in the field, the paper seeks to elucidate the underlying principles, techniques, applications, challenges, and recent developments within each of these paradigms. This comprehensive analysis is intended to deepen the understanding of how these systems operate and how they can be utilized.

4. Extractive Question Answering (EQA)

4.1. Definition and characteristics of EQA systems

Definition 1. EQA systems are a type of QA system that retrieve answers directly from a given corpus of text or data (Wang et al., 2022; Assem et al., 2021).

EQA involves working with a given text passage T and a question Q, as illustrated in Fig. 2. The process begins by taking the input pair (T, Q) as outlined in algorithm 1. The next step is to identify and locate the span $S \subseteq T$ that answers the question Q. Finally, the relevant span S is extracted and returned as the answer A, which equals S. Mathematically, EQA can be represented as:

$$\text{EQA}(T, Q) = \arg\max_{S \subseteq T} P(S \mid T, Q)$$

where $P(S \mid T, Q)$ denotes the probability of span S being the correct answer given T and Q.

Algorithm 1 EQA

Input: Text passage T, Question Q
Output: Answer span S

```
max_prob ← 0
S ← empty span
for each possible span s in T do
    prob ← P(s | T, Q)
    if prob > max_prob then
        max_prob ← prob
        S ← s
    end if
end for
return S
```

Overall, EQA iterates through all possible spans in the text passage to find the one with the highest probability of being the correct answer.

For instance, let us consider the text passage below along with a corresponding question:

T: Albert Einstein was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern physics. His work is also known for its influence on the philosophy of science. Einstein is best known in popular culture for his mass–energy equivalence formula $E = mc^2$. He received the 1921 Nobel

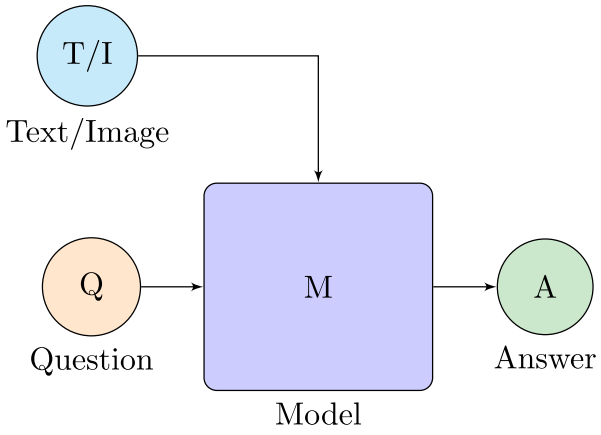


Fig. 2. Abstract representation of QA systems: In this context, the predictive model (M) within a QA system could vary, incorporating options such as extractive, generative, or VQA, depending on the nature of the input and the resultant output. The input for the QA task might consist of a question (Q) paired with either a text passage or an image, while the resulting answer (A) could manifest as extractive response directly drawn from the input, or a generative response based on the input's content.

Prize in Physics for his "services to theoretical physics", in particular, his discovery of the law of the photoelectric effect, a pivotal step in the development of quantum theory.

Q: For what discovery did Albert Einstein receive the Nobel Prize in Physics in 1921?

To implement EQA, the process involves two main steps. First, identification: locating the span within the text passage (T) that directly addresses the question (Q). In the given example, the answer to the question is explicitly provided in the passage: "He received the 1921 Nobel Prize in Physics for his'services to theoretical physics', in particular, his discovery of the law of the photoelectric effect..." Second, extraction: once the relevant span is identified, it is extracted. In this case, the identified span is: "his discovery of the law of the photoelectric effect". Therefore, for this specific example, the output of EQA would be: "his discovery of the law of the photoelectric effect".

Moreover, EQA systems typically consist of several key characteristics:

- Retrieval-Based Approach: EQA systems employ a retrieval-based approach, where answers are extracted directly from a pre-existing collection of documents or text passages (Chen and Zulkernine, 2021; Demner-Fushman and Lin, 2006). The system does not generate new answers but instead identifies and extracts relevant information from the provided sources.
- Keyword Matching: EQA systems often rely on keyword matching techniques to identify relevant passages or documents containing the answer to a given question. They search for specific keywords or phrases within the corpus and extract sentences or paragraphs that contain these keywords (Zhu et al., 2021).
- Limited Understanding: Unlike GQA systems, EQA systems have limited understanding of the context or semantics of the text (Hu et al., 2023; Lee et al., 2016). They primarily focus on matching keywords or patterns without deeper comprehension of the underlying meaning of the text (Butt et al., 2021).
- Document Retrieval: EQA systems may utilize document retrieval techniques (Zhu et al., 2021) to identify a set of candidate documents or passages that are likely to contain the answer to the given question (Singh et al., 2021b). These documents are then further processed to extract the most relevant information.
- Scalability: EQA systems need to be scalable to handle large volumes of text or documents efficiently (Seo et al., 2018; Feng et al., 2020). They should be able to retrieve answers quickly and accurately from a vast corpus of information.

Overall, EQA systems provide a straightforward approach to QA by extracting answers directly from existing text sources. While they may lack the ability to generate new answers or understand complex contexts, they are effective for tasks that involve retrieving factual information (Wang et al., 2020a) from structured or unstructured data.

4.2. Methodologies

The methodologies employed by EQA systems typically involve passage retrieval, ranking, and selection techniques to identify and extract relevant information from a corpus of text or documents:

- Passage Retrieval: Passage retrieval (Tellex et al., 2003; Zhu et al., 2021) is the initial step in EQA systems, where a set of candidate passages or documents likely to contain the answer to the given question is identified. This process involves techniques such as keyword matching, document indexing, and inverted indexing. Keyword matching (Demner-Fushman and Lin, 2006; Fader et al., 2014; Zhu et al., 2021) involves searching for specific keywords or phrases in the corpus that match those present in the question. Document indexing (Tsatsaronis et al., 2015) creates an index of the entire corpus, allowing for efficient retrieval of relevant documents based on keyword queries. Inverted indexing (Qu et al., 2020), on the other hand, indexes each word or term to the documents in which it appears, enabling fast retrieval of documents containing specific keywords. In question-answering systems, information retrieval is pivotal in identifying potential documents or passages that could hold the answer to posed questions. Popular IR algorithms encompass Boolean retrieval (Lee et al., 2006; Rakotoson et al., 2022), vector space models (Kolomiyets and Moens, 2011), and probabilistic approaches such as BM25 (Robertson et al., 2009).
- Ranking: Once candidate passages or documents are retrieved, the next step is to rank them based on their relevance to the question (Zhu et al., 2021; Iyer et al., 2021). Ranking methodologies may include techniques such as term frequency-inverse document frequency (TF-IDF), BM25, and neural network-based approaches. TF-IDF (Arroyo-Fernández et al., 2019) calculates the importance of a term in a document relative to its frequency in the entire corpus, while BM25 (Robertson et al., 2009) considers factors such as term frequency and document length to rank documents. Neural network-based ranking models (Nguyen and Le, 2018) leverage deep learning techniques to learn complex patterns and relationships between questions and passages, leading to more accurate rankings (Lukovnikov et al., 2017; Abbasiantaeb and Momtazi, 2021). However, retrieval techniques such as TF-IDF and BM25 utilize sparse representations. While these sparse techniques provide lightweight and efficient solutions (Ahmed et al., 2022; Mao et al., 2020), they fall short in semantic matching ability and face challenges in retrieving relevant passages when there is minimal lexical overlap.
- Selection: After ranking the candidate passages, the final step is to select the most relevant passage or passages as the answer to the question (Tellex et al., 2003; Cui et al., 2005). Selection methodologies may include threshold-based approaches (Zhu et al., 2021), where passages above a certain relevance score threshold are selected, or ML-based approaches (Yen et al., 2013; Tahsin Mayeesha et al., 2021), where a classifier is trained to predict the relevance of passages to the question. Ensemble methods (Cheng et al., 2021; Moholkar and Patil, 2021) combining multiple ranking and selection techniques may also be employed to improve the overall performance of EQA systems.
- Natural Language Processing: NLP comprises a spectrum of methodologies aimed at comprehending and handling natural language text. These techniques are instrumental in dissecting the structure (Olivetti et al., 2020), syntax (Srihari et al., 1999),

semantics (Berant et al., 2013), and contextual nuances embedded within textual data (Louis et al., 2024). Within QA frameworks, NLP plays a pivotal role in deciphering user queries and extracting pertinent information from diverse text sources. The array of NLP tasks pertinent to QA encompasses tokenization (Ly et al., 2020; Fares et al., 2013), part-of-speech tagging (Chiche and Yitagesu, 2022), named entity recognition (Nadeau and Sekine, 2007), syntactic parsing (Manning and Schutze, 1999), semantic analysis (Maulud et al., 2021), and discourse analysis (Liakata et al., 2013). Moreover, cutting-edge NLP models, exemplified by deep learning architectures like Transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2018), LLaMA (Touvron et al., 2023), and GPT-3 (Floridi and Chiriatti, 2020) have demonstrated considerable advancements across various NLP domains, including question-answering.

However, Integration of IR and NLP techniques in question-answering systems (Gaizauskas and Humphreys, 2000; Abacha and Zweigenbaum, 2015) allows for efficient retrieval and processing of relevant information to produce accurate answers. These techniques enable systems to handle a wide range of questions and text sources effectively, contributing to the development of robust and versatile question-answering systems. Overall, the methodologies of passage retrieval, ranking, and selection play a crucial role in the effectiveness and accuracy of EQA systems by enabling them to efficiently identify and extract relevant information from large corpora of text or documents.

4.3. Applications

EQA systems find various applications across different domains, primarily focusing on tasks related to IR, factoid QA, and reading comprehension. Here's an overview of these applications:

- Information Retrieval: EQA systems play a crucial role in IR tasks (Manna et al., 2020; Chen and Zulkernine, 2021; Demner-Fushman and Lin, 2006; Tellex et al., 2003; Zhu et al., 2021; Demner-Fushman and Lin, 2006; Fader et al., 2014; Zhu et al., 2021), where users seek specific pieces of information from a vast corpus of text or documents. These systems enable users to pose natural language questions and retrieve relevant documents or passages containing the desired information. IR applications include search engines (Kwok et al., 2001; Ren et al., 2021), document retrieval systems (Zhu et al., 2021; Singh et al., 2021b), and knowledge base querying (Liu et al., 2020a; Das et al., 2017).
- Factoid Question Answering: In factoid question answering, EQA systems are employed to provide concise and factual answers to questions posed by users (Er and Cicekli, 2013; Soricut and Brill, 2006). These questions typically require specific information, such as definitions (Saggion, 2004), dates (Raiman and Miller, 2017), locations (Xu et al., 2021), or numerical data (Pearce et al., 2021). EQA systems retrieve relevant passages or documents containing the answer and present it to the user in a structured format. Factoid QA applications include virtual assistants (Reddy et al., 2019), chatbots (Alruqi and Alzahrani, 2023), and information kiosks (Saha et al., 2018; Dhingra et al., 2017).
- Reading Comprehension: EQA systems are utilized in reading comprehension tasks to assess a system's ability to understand and answer questions based on a given text passage or document (Zhao et al., 2021; Zhu et al., 2021). These systems are trained on datasets containing question–answer pairs and corresponding text passages such as Jaquad (So et al., 2022), SQuAD (Rajpurkar et al., 2016), for additional datasets on reading comprehension, refer to Bai and Wang (2021), enabling them to extract answers directly from the text. Reading comprehension applications include educational tools, language tutoring systems, and assessment platforms.

Overall, EQA systems have diverse applications in tasks requiring the retrieval of factual information from text sources. They enable efficient IR, facilitate fact-based QA, and support reading comprehension tasks in various domains ranging from education and research to information access and decision support.

4.4. Evaluation scores and datasets

Evaluation scores play a crucial role in assessing the performance of QA systems. They provide quantitative measures to evaluate the accuracy, effectiveness, and overall performance of these systems. Three commonly used QA datasets for evaluating QA systems are Stanford Question Answering Dataset SQuAD (Rajpurkar et al., 2016), TREC QA corpus (Voorhees and Tice, 2000), and MS-MARCO (Nguyen et al., 2016).

1. Evaluation Scores: Evaluation of EQA systems often involves scores such as precision, recall, and F1-score, which measure the system's ability to retrieve relevant answers accurately and completely (Lewis et al., 2019b; Farea et al., 2022; Farea and Emmert-Streib, 2024)
2. Stanford Question Answering Dataset (SQuAD): SQuAD (Rajpurkar et al., 2016) is a widely used dataset for machine reading comprehension tasks. It consists of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text (span) from the corresponding passage. The primary evaluation scores used for SQuAD are Exact Match (EM) and F1 score. EM measures the percentage of predicted answers that match exactly with the ground truth answers. F1 score is the harmonic mean of precision and recall, where precision is the ratio of the number of correct words in the predicted answer to the total number of words, and recall is the ratio of the number of correct words to the total number of words in the ground truth answer.
3. Text REtrieval Conference Question Answering (TREC QA): TREC QA (Voorhees and Tice, 2000) is a collection of datasets used in the TREC competitions for evaluating QA systems. These datasets contain various types of questions, including fact-based, definition, and list questions, along with corresponding answer passages. The evaluation score used for TREC QA varies depending on the specific task and dataset. Common scores include Mean Reciprocal Rank (MRR). MRR measures the average of the reciprocal ranks of the correct answers.
4. Microsoft MAchine Reading COmprehension (MS MARCO): MS MARCO (Nguyen et al., 2016) is a large-scale dataset for machine reading comprehension and question answering. It consists of real user queries from the Bing search engine along with corresponding documents. The primary assessment criteria employed for MS MARCO focus on accuracy, precision, and recall to gauge the precision of numerical responses. Additionally, scores such as ROUGE-L (Lin, 2004) and the phrasing-aware evaluation framework (Mitra et al., 2016) are utilized for evaluating lengthy textual answers. The phrasing-aware evaluation framework is designed to address the nuances of natural language diversity when assessing extended textual responses. This evaluation process necessitates a substantial quantity of reference answers per query, each meticulously curated by distinct human editors. Consequently, it offers a natural means to gauge the variety of ways individuals may articulate responses to the same question. A range of pairwise similarity-based scores is then applied to incorporate consensus among diverse reference answers for evaluation purposes. These scores, simple adaptations of established measures such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), have demonstrated enhanced correlation with human evaluations. Consequently, in their experimental endeavors, a subset of MS MARCO featuring

multiple answers per query was employed to assess model performance using both BLEU and pa-BLEU (Mitra et al., 2016) as evaluation scores.

These datasets and evaluation scores (Bai and Wang, 2021) provide standardized benchmarks for assessing the performance of QA systems across different domains and tasks. They enable researchers to compare the effectiveness of various approaches and techniques in the field of QA.

4.5. Challenges and limitations

- Ambiguity Resolution: One of the primary challenges faced by EQA systems is ambiguity resolution (Allam and Haggag, 2012). Natural language is inherently ambiguous (Kamsties and Peach, 2000; Yang et al., 2011), and questions may have multiple valid interpretations (Palangi et al., 2018; Cohen et al., 2014). EQA systems often struggle to disambiguate between different meanings of words or phrases (Towell and Voorhees, 1998; Liu et al., 2005), leading to inaccurate or irrelevant answers (Perera et al., 2020). Resolving ambiguity requires deeper understanding of context (Guo et al., 2021), which is often lacking in retrieval-based approaches (Keyvan and Huang, 2022).
- Entity Recognition: EQA systems rely heavily on entity recognition to identify relevant entities mentioned in the text. However, entity recognition can be challenging, especially in domains with complex or ambiguous entity names (Li et al., 2017). EQA systems may struggle to accurately identify entities, leading to incorrect answers or missed information (Hirschman and Gaizauskas, 2001).
- Scalability: Scalability (Assem et al., 2021; Cui et al., 2019) is another significant challenge for EQA systems, particularly when dealing with large volumes of text or data. As the size of the corpus increases, the time and computational resources required for document retrieval and answer extraction also increase. Scaling EQA systems to handle large datasets efficiently without sacrificing performance is a major technical challenge.
- Language Understanding: EQA systems often lack deeper language understanding capabilities (Clark et al., 2020), relying primarily on keyword matching techniques. This limits their ability to comprehend complex language structures, nuances, and linguistic variations. As a result, EQA systems may struggle with questions that require deeper semantic understanding or reasoning (Hovy et al., 2000).
- Answer Consistency: Another limitation of EQA systems is the lack of consistency in answers (Ribeiro et al., 2019; Honovich et al., 2021). Since answers are extracted directly from the text corpus, they may vary depending on the specific passages or documents retrieved by the system. Inconsistencies in answers can erode user trust and confidence in the system.
- Domain Specificity: EQA systems may perform well in certain domains where the language is structured and unambiguous (Ma et al., 2015; Sheker et al., 2016), but they may struggle in more specialized or domain-specific contexts. Adapting EQA systems to different domains requires extensive training data and domain-specific knowledge, which may not always be available.

Addressing these challenges and limitations requires advancements in natural language understanding, entity recognition, and document retrieval techniques. Future research efforts in these areas aim to improve the performance and scalability of EQA systems, enabling them to provide more accurate and reliable answers across a wide range of domains and use cases.

4.6. Recent advancements and research trends

Recent advancements and research trends in EQA systems have focused on improving the accuracy, efficiency, and versatility of these systems. Some notable advancements include:

- Deep Learning Techniques: Researchers have increasingly utilized deep learning models, such as RNNs (Serban et al., 2016a), CNNs (Li et al., 2021), and transformer-based architectures (Kumari and Pushphavati, 2022) like BERT (Devlin et al., 2018), LLaMA (Touvron et al., 2023) and GPT (Klein and Nabi, 2019), to enhance the performance of EQA systems. These models leverage large-scale pretraining on text corpora to better capture semantic relationships and contextual information, resulting in improved answer extraction accuracy.
- Transfer Learning: Transfer learning techniques have been applied to EQA systems to leverage knowledge from pretrained models and adapt them to specific QA tasks (Chung et al., 2017; Min et al., 2017). By fine-tuning pretrained language models on domain-specific datasets or adjusting them for different languages, researchers have achieved significant improvements in EQA performance without the need for extensive labeled data.
- Multimodal EQA: With the increasing availability of multimedia data (Singh et al., 2021a; Ben-Younes et al., 2017), there has been growing interest in multimodal EQA systems that can extract answers from both text and other modalities such as images, videos, or audio. These systems integrate information from multiple modalities to provide more comprehensive and accurate answers, opening up new avenues for applications in fields like multimedia analysis, robotics, and virtual assistants.
- Efficient Retrieval Techniques: To address the scalability challenges associated with large text corpora, researchers have developed efficient retrieval techniques for EQA systems (Gupta and Gupta, 2012). These techniques leverage indexing (Seo et al., 2019; Tsatsaronis et al., 2015), caching (Cao et al., 2020), and approximate search algorithms to quickly identify relevant passages or documents containing the answer to a given question (Castel et al., 2021), enabling real-time or near-real-time QA over vast datasets (Kim et al., 2020).
- Domain-Specific EQA: Tailoring EQA systems to specific domains or verticals, such as healthcare (Chen and Zulkernine, 2021; Chang et al., 2023), finance (Nararatwong et al., 2022), or legal (Yadav and Singh, 2024), has gained attention in recent years. Domain-specific EQA models are trained on specialized datasets and fine-tuned to understand domain-specific terminology and context, resulting in more accurate and relevant answers for users in those domains.
- Evaluation and Benchmarking: The development of standardized evaluation benchmarks and datasets has played a crucial role in advancing EQA research. Benchmark datasets, such as SQuAD and MS MARCO, provide standardized tasks and evaluation scores for comparing the performance of different EQA models, facilitating fair and rigorous evaluation of system performance.

Overall, recent advancements and research trends in EQA systems are driving progress towards more accurate, efficient, and domain-adaptive QA solutions that can effectively extract answers from large volumes of text and multimedia data across diverse domains and applications.

5. Generative Question Answering (GQA)

5.1. Definition and characteristics of GQA systems

Definition 2. GQA systems represent a type of QA system that constructs answers using NLG techniques (Lewis and Fan, 2018), rather than merely retrieving them from pre-existing text sources.

GQA involves the process of generating answers to questions based on a given text passage and question, as depicted in Fig. 2. This process begins with the input of the text passage T and question Q, which are then encoded into a latent and meaningful representation, as outlined in algorithm 2. Subsequently, this representation is decoded to produce the answer A. This process can be mathematically represented as:

$$\text{GQA}(T, Q) = \arg\max_{A} P(A|T, Q)$$

where GQA seeks to find the answer A that maximizes the probability $P(A|T, Q)$, which represents the likelihood of A being the correct answer given T and Q.

Algorithm 2 GQA(T, Q)

Input: Text passage T, Question Q
Output: Generated answer A

```
R ← Encode(T, Q)
A ← empty sequence
while ¬ IsComplete(A) do
    a_i ← GenerateNextToken(R, A)
    Append a_i to A
end while
return A
```

Overall, GQA encodes the text passage and question into a latent representation and then generates the answer token by token based on this representation.

To illustrate, let us say we're presented with the subsequent text passage and question:

T: Mount Everest, at 8848.86 m (29,031.7 ft), is the world's highest mountain.

Q: How tall is Mount Everest?

Now, instead of straightforwardly generating an extractive answer, the GQA model could generate a response by understanding the context of the question and providing a more descriptive or explanatory answer, which could still contain the relevant information but in a paraphrased form. In this scenario, the GQA model might generate a more dynamic answer, such as:

A: Mount Everest stands proudly as the tallest peak on Earth, soaring to an impressive height of 8848.86 m (29,031.7 ft).

This answer still provides the height of Mount Everest but presents it in a more descriptive and informative manner, avoiding direct extraction from the text passage.

Here are some defining characteristics of GQA systems:

- Generation-Based Approach: GQA systems employ a generation-based approach (Bidgoly et al., 2022; Kurdi et al., 2020), where answers are generated from scratch based on the understanding of the question and the relevant context. Unlike EQA, GQA systems create new responses rather than selecting them from a fixed set of options.
- Understanding of Context: GQA systems demonstrate a deeper understanding of the context and semantics of the text (Lewis and Fan, 2018; Liu et al., 2019). They analyze the meaning of the question and consider the broader context in order to generate accurate and relevant answers that may not be explicitly stated in the input data.
- Natural Language Processing: GQA systems heavily rely on NLP techniques (Yin et al., 2015) to comprehend the nuances of the text, including syntactic structures, semantics, and discourse. These systems utilize various NLP models, such as neural language models, to generate fluent and coherent responses.
- Creativity and Flexibility: GQA systems exhibit creativity and flexibility in generating diverse answers to a wide range of questions (Liu et al., 2020b). They can produce responses that go beyond simple factual information (Soricut and Brill, 2006), incorporating reasoning, inference, and abstraction to provide more insightful and nuanced answers (Yin et al., 2015).
- Data-driven Learning: GQA systems leverage large-scale datasets and ML algorithms to learn patterns and relationships from textual data (Wang et al., 2018; Liu et al., 2019). They are trained on extensive corpora of text and use statistical models to predict and generate responses based on input questions and contexts.

In summary, GQA systems represent a sophisticated approach to QA, leveraging advanced NLP techniques to generate contextually relevant and coherent answers. They offer greater flexibility and expressiveness compared to EQA, making them suitable for a wide range of applications in natural language understanding and generation.

5.2. Methodologies

Methodologies such as NLG, Sequence-to-Sequence (Seq2Seq), and related deep learning modeling are pivotal in advancing the capabilities of GQA systems, particularly in the realm of understanding and generating human-like responses:

- Natural Language Generation encompasses techniques that enable machines to produce coherent and contextually relevant natural language text (Zhang et al., 2021). In the context of GQA systems, NLG methodologies facilitate the generation of answers to questions posed in natural language. These methodologies often involve the use of linguistic rules (Yin et al., 2015), templates (Zhang et al., 2021), and ML algorithms to transform structured data or knowledge into human-readable text (Yin et al., 2015). NLG techniques ensure that the generated responses are grammatically correct, semantically meaningful, and contextually appropriate.
- Sequence-to-Sequence (Seq2Seq) modeling, on the other hand, is a deep learning architecture commonly used for tasks involving sequence generation (Song et al., 2017; Yin et al., 2015), such as machine translation (Lewis et al., 2019a), text summarization (Savery et al., 2020), and question answering (Gunasekara et al., 2021). In Seq2Seq models, an encoder–decoder framework (Yin et al., 2015) is employed, where an encoder neural network processes the input sequence (e.g., a question), and a decoder network generates the output sequence (e.g., an answer). This methodology has been adapted for GQA systems, where the encoder processes the question, and the decoder generates the corresponding answer. The model learns to map input questions to output answers by capturing complex patterns and dependencies within the data.
  Moreover, GQA systems employ various techniques to understand and generate answers to questions. Three prominent techniques utilized in GQA systems are RNNs, transformers, and attention mechanisms, as outlined below:
- Recurrent Neural Networks: RNNs are a class of neural networks particularly effective for sequential data processing, making them well-suited for GQA tasks. They process input sequences step-by-step, maintaining an internal state that captures information from previous steps. This ability to capture temporal dependencies enables RNNs to model sequential data effectively (Lewis and Fan, 2018), making them suitable for tasks such as language modeling and sequence generation in GQA systems. However, traditional RNNs suffer from issues such as vanishing gradients and difficulty in capturing long-range dependencies (Tang et al., 2017), leading to the development of more advanced architectures like Long Short-Term Memory (LSTM) (Lewis and Fan, 2018) and Gated Recurrent Unit (GRU) (Shen et al., 2021) to address these shortcomings.

- Transformers: Transformers have emerged as a powerful architecture for various NLP tasks, including GQA. Unlike RNNs, transformers process input data in parallel, allowing for more efficient computation and capturing long-range dependencies effectively. The transformer architecture consists of self-attention mechanisms (Vaswani et al., 2017) that enable it to weigh the importance of different input tokens when generating output representations. This attention mechanism allows transformers to focus on relevant parts of the input sequence, making them well-suited for tasks requiring context understanding, such as language translation and question answering.
- Attention Mechanisms: Attention mechanisms (Vaswani et al., 2017) play a crucial role in both RNNs and transformer architectures. They enable models to selectively focus on relevant parts of the input sequence when generating output representations. In GQA systems, attention mechanisms allow the model to attend to specific words or tokens in the question and relevant context in the input data, facilitating accurate answer generation. By assigning higher weights to informative tokens and suppressing irrelevant ones, attention mechanisms improve the model's ability to understand and generate answers to complex questions.
- Reinforcement Learning from Human Feedback (RLHF): is a method that enhances ML models by incorporating human input into their training process (Bai et al., 2022; Zhu et al., 2023). RLHF plays a crucial role in refining the quality and relevance of generated answers. Initially, a model is trained using traditional reinforcement learning techniques on a broad dataset to learn basic behaviors. Human feedback is then collected, which might involve ranking outputs, providing corrections, or offering qualitative assessments. This feedback is used to develop a reward model that reflects human preferences and values. The reinforcement learning algorithm subsequently fine-tunes the model based on the rewards predicted by this reward model, aligning its behavior more closely with human expectations. This iterative process helps improve the model's performance and adaptability, ensuring that it produces more accurate and contextually appropriate outputs. By integrating human insights, RLHF enables models to better align with user values and preferences, enhancing their effectiveness and overall user experience.

Over all, RNNs, transformers, and attention mechanisms are fundamental techniques used in GQA systems to process input data, capture contextual information, and generate accurate answers to questions. Each of these techniques has its strengths and weaknesses, and their effectiveness depends on the specific requirements of the GQA task and the characteristics of the input data.

The integration of NLG techniques and Seq2Seq modeling in GQA systems enables them to generate responses that are not only contextually relevant but also syntactically and semantically coherent. By leveraging these methodologies, GQA systems can produce human-like answers to a wide range of questions, including those that require complex reasoning or inference. However, these systems might face significant challenges when dealing with complex or ambiguous queries. For instance, they may struggle to correctly interpret the context or meaning of a question when it is not clearly defined. This could lead to errors in answering. For example, if a user asks, 'Can you tell me how to change a tire in my car?' without specifying the make or model, the system might struggle to provide an accurate response, as different cars require different methods. Similarly, when faced with ambiguous queries that involve nuanced reasoning — such as 'What should I do if I'm feeling overwhelmed at work?' — the system might offer generic advice that misses the specific needs of the user. In dynamic or uncertain situations, such as real-time problem-solving, the system might not always be able to adapt its answers quickly enough to remain accurate.

Furthermore, ongoing research and advancements in NLG and Seq2Seq modeling continue to enhance the capabilities of GQA systems (Yao et al., 2024; Ji et al., 2023), making them increasingly adept at understanding and generating natural language responses. Recent advancements, including the release of models like ChatGPT-4 (2003) (Spennemann, 2023), PaLM 2 (2023) (Anil et al., 2023), and Google's Gemini (2024) (Team et al., 2023; Imran and Almusharraf, 2024), which offer enhanced reasoning abilities, better context management, and greater adaptability, are expanding the capabilities of these systems.

5.3. Applications

GQA systems are utilized across various domains (Izacard and Grave, 2020), including conversational agents (Li et al., 2023), customer support systems (Hardalov et al., 2018), virtual assistants (Reddy et al., 2019), and educational platforms (Chheang et al., 2024). They excel at producing human-like responses to open-ended queries and facilitating natural language conversations. The applications of these systems can be summarized as follows:

- Open-Domain Question Answering: GQA systems are employed in open-domain QA tasks where users can pose questions on any topic (Izacard and Grave, 2020), and the system retrieves relevant information from large knowledge bases or text corpora. These systems enable users to find answers to factual questions, explore complex topics, and collect information from various sources. Most importantly, they do all of this without any predefined limitations.
- Dialogue Systems: GQA models serve as crucial components in dialogue systems (Serban et al., 2016b), enabling natural and informative interactions between users and machines. In conversational settings, such as chatbots or virtual assistants, GQA systems help to provide accurate and relevant responses to user queries, contributing to seamless and engaging conversational experiences.
- Conversational Agents: GQA technology is integral to the development of conversational agents or virtual assistants deployed in various applications (Ramesh et al., 2017), including customer service, IR, and personal assistance. By leveraging GQA capabilities, conversational agents can understand user inquiries, retrieve pertinent information from vast knowledge repositories, and deliver contextually appropriate responses, thereby enhancing user satisfaction and productivity.
  Dialogue systems and conversational agents are closely related but differ in their scope and complexity. Dialogue systems are designed for structured interactions, managing conversations through predefined rules and patterns. They are often used for specific tasks, such as customer service bots or automated booking systems, focusing on delivering targeted information efficiently. In contrast, conversational agents are broader and aim for a more natural and flexible interaction. They leverage advanced NLP and ML to engage in more human-like conversations, capable of handling a wide range of topics and learning from interactions. While dialogue systems follow a more rigid structure, conversational agents offer dynamic and adaptable communication, providing a richer user experience.
- Information Retrieval: GQA systems facilitate efficient IR by enabling users to pose natural language queries (Yin et al., 2015) and receive precise answers from structured or unstructured data sources. These systems are utilized in search engines, digital libraries, and knowledge management platforms to streamline the process of accessing relevant information and knowledge resources.

- Educational Platforms: GQA technology is integrated into educational platforms and e-learning environments to support learners in acquiring knowledge, clarifying concepts, and resolving queries (Chheang et al., 2024). By providing instant access to accurate information and explanations, GQA systems enhance the effectiveness of online learning experiences and foster knowledge acquisition across diverse educational domains.
- Decision Support Systems: GQA models contribute to decision support systems by assisting users in obtaining timely and relevant information for decision-making processes (Hardalov et al., 2018). In domains such as healthcare, finance, and business intelligence, GQA systems aid users in accessing critical insights, analyzing data, and making informed decisions based on comprehensive knowledge and evidence.

Overall, the applications of GQA extend across various domains and play a pivotal role in enhancing information access, communication, and decision-making processes in both individual and organizational contexts.

5.4. Evaluation scores and datasets

The evaluation scores and datasets associated with GQA encompass a range of measures and datasets tailored to assess the performance of GQA systems. Here is an overview of some key evaluation scores and datasets:

1. Evaluation Scores:

    - Evaluation Scores: While F1 is a practical metric for Evaluative EQA, its effectiveness can vary depending on the specific QA task due to different types of answers. Conventional metrics such as CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and F1 are often adequate for common QA datasets (Chen et al., 2019). However, these metrics may limit the complexity of QA datasets that can be effectively evaluated, particularly in free-form QA scenarios where models can generate complex and abstract responses. Therefore, there is a need for new evaluation metrics that go beyond n-gram-based comparisons. Metrics such as BERTScore (Zhang et al., 2019), KPQA (Lee et al., 2020), and others like perplexity (Klakow and Peters, 2002; Xiong et al., 2016), which evaluate the uncertainty or unpredictability, fluency (Van Der Lee et al., 2019), and relevance (Zhong et al., 2022) of generated answers compared to human references, are becoming increasingly important.
    - Conversational Question Answering(CoQA): CoQA (Reddy et al., 2019) is a dataset designed for evaluating conversational QA systems. It consists of questions posed by humans based on a provided passage, along with answers generated by other humans. The dataset contains diverse question types and requires systems to understand and generate coherent responses in a conversational context.

These evaluation scores and datasets play a crucial role in assessing the performance and capabilities of GQA systems. They provide standardized measures and benchmark datasets that enable researchers to compare different approaches, track progress, and identify areas for improvement in the field of question answering.

5.5. Challenges and limitations

The GQA paradigm, while promising, encounters several challenges and limitations (Yin et al., 2015; Pandya and Bhatt, 2021), primarily revolving around context understanding, coherence, and data scarcity.

- Context Understanding: GQA systems face challenges in comprehensively understanding the context of questions and documents (Song et al., 2017; Yin et al., 2015). Unlike simple keyword matching approaches, GQA requires a deeper understanding of the semantic meaning and context of the text to generate accurate and coherent answers. Achieving this level of context understanding remains a significant challenge, especially in tasks involving complex or ambiguous queries.
- Coherence: Generating coherent and contextually appropriate answers (Wang et al., 2019) is another challenge for GQA systems. While the models may be adept at producing grammatically correct responses, ensuring that the generated answers are contextually relevant and coherent with the given question and surrounding text remains a significant hurdle. Maintaining coherence across longer responses or in scenarios with multiple possible interpretations adds to the complexity.
- Data Scarcity: GQA systems often require large amounts of high-quality training data to learn effectively. However, obtaining such data, particularly annotated datasets that capture the nuances of language and context, can be challenging and expensive. Data scarcity (Sufi, 2024) can limit the performance and generalization capabilities of GQA models, particularly in domains or languages with limited resources.
- Hallucinations: Large Language Models (LLMs) possess impressive capabilities but also face problems like hallucinations, slow knowledge updates, and lack of transparency. Occasionally, they give answers that are wrong, unrelated, misleading, or completely made up, even though they might sound believable (Sriramanan et al., 2024; Galitsky, 2025). These problems happen because LLMs rely on finding patterns instead of truly understanding the information they are working with Huang et al. (2023) and Zhang et al. (2023). For example, LLMs store knowledge in their model parameters, which come from big datasets. However, these models face some issues. First, they have difficulty keeping rare or very specific knowledge. Second, because their parameters cannot be updated quickly, the knowledge they have becomes old. Also, if the model's parameters are made larger to improve performance, the cost of training and using it also increases. If an LLM is trained on a small or non-diverse dataset, it might make mistakes or give answers that do not match the real content. The model might also miss important details, making its answers not capture all the subtleties of a question or image (like in VQA systems). These problems make things even worse. Hallucinations can spread false information and make people trust LLMs less, especially in important or sensitive situations.
  A good solution for these challenges is *Retrieval-Augmented Generation (RAG)*. This approach combines a parameterized model with an external database that is not based on parameters, allowing the model to get updated information from this database. This makes LLMs more accurate and efficient without needing constant updates to the model's parameters.
  RAG combines the best parts of both retrieval-based and generative models, ensuring that the answers the model gives are accurate and based on real-world data (Lewis et al., 2020). In RAG, the model first finds relevant information from a large external knowledge source before giving an answer (Gao et al., 2023). This step helps ensure that the answers are based on real, factual sources, reducing the risk of hallucinations.
  RAG also increases trust by citing its sources and allowing for regular knowledge updates, which include specific information for certain fields. It combines the knowledge stored in LLMs with external databases, making it an important method for using LLMs. Additionally, RAG helps reduce hallucinations by preventing the model from focusing too much on its training data and by using a larger variety of knowledge sources.

In addition to RAG, other methods are being explored to reduce hallucinations in generative models. These include fact-checking models that compare answers with reliable sources, using higher-quality training data, and applying controlled text generation methods that limit the model's responses. Also, techniques like RLHF, where human input is used to improve the model's outputs over time, have shown promise in improving accuracy (Augenstein et al., 2024; Atanasova, 2024). To sum up, even though hallucinations are still a big challenge in GQA systems, combining external retrieval methods with other advanced techniques brings hope for building more accurate, reliable, and trustworthy AI models.

Addressing these challenges requires advancements in natural language understanding, model architectures, and training methodologies. Techniques such as pre-training on large-scale datasets, fine-tuning on domain-specific corpora, and incorporating external knowledge sources can help improve context understanding and coherence in GQA systems. Additionally, research efforts focused on data augmentation, transfer learning, and semi-supervised learning can mitigate the impact of data scarcity, enabling GQA models to perform more effectively across a range of tasks and domains.

5.6. Recent advancements and research trends

Recent advancements and research trends in GQA systems have seen significant progress, particularly in enhancing the ability of models to understand and generate human-like responses to questions posed in natural language. Some notable advancements and research trends with GQA include:

- Deep Learning Architectures: There has been a growing focus on developing more sophisticated deep learning architectures for GQA systems. Models such as transformers, particularly variants like BERT (Qu et al., 2019), LLaMA (Touvron et al., 2023) and GPT (Klein and Nabi, 2019), have shown remarkable performance in understanding and generating text, including answers to questions.
- Pre-training and Fine-tuning: Pre-training large language models on vast amounts of text data followed by fine-tuning on specific QA tasks has become a common practice (Chang et al., 2023). This approach has led to significant improvements in the performance of GQA systems by leveraging the vast amount of linguistic knowledge encoded in pre-trained models.
- Multi-hop Reasoning: Recent research has focused on enhancing the reasoning capabilities of GQA systems, particularly in scenarios that require multi-hop reasoning (Liu et al., 2020a). These advancements enable models to integrate information from multiple passages or documents to derive answers that require complex inference and reasoning.
- Semantic Understanding: Efforts have been made to enhance the semantic understanding capabilities of GQA systems (Wang et al., 2019). This includes incorporating techniques from NLU to enable models to grasp the nuanced meanings of questions and provide more accurate and contextually relevant answers (Liu et al., 2019; Niu et al., 2021).
- Multimodal Integration: Another emerging trend is the integration of multimodal information (Bang et al., 2023; Ben-Younes et al., 2017; Singh et al., 2021a), such as combining text with images or other types of data. Multimodal GQA systems aim to answer questions that require understanding both textual and visual information, leading to more comprehensive and informative responses.
- Evaluation and Benchmarking: There is ongoing research into developing standardized evaluation benchmarks and scores for assessing the performance of GQA systems. This includes creating datasets with diverse question types, linguistic complexities, and reasoning requirements to provide a comprehensive evaluation of model capabilities (Chen et al., 2019; Yin et al., 2015).
- Ethical and Fairness Considerations: As GQA systems become more sophisticated and widely deployed, there is increasing awareness of ethical considerations and biases inherent in these systems (Hua et al., 2024). Research efforts are underway to address issues related to fairness, transparency, and accountability in GQA models to ensure equitable outcomes and mitigate potential biases.

Overall, recent advancements and research trends with GQA systems have been transformative, paving the way for more robust, intelligent, and human-like QA capabilities across various domains and applications.

6. Visual question answering (VQA)

6.1. Definition and characteristics of VQA systems

Definition 3. VQA systems represent a fusion of computer vision and NLP techniques, enabling machines to comprehend and respond to questions about visual content, such as images or videos.

VQA involves the analysis of an image (I) along with a posed question (Q), as illustrated in Fig. 2. The process starts with the input of both the image and question (I, Q), as outlined in algorithm 3. The image is encoded into a visual representation (V), which is then combined with the question. Subsequently, this combined representation is decoded to derive an answer (A). This process is mathematically represented as:

$$\text{VQA}(I, Q) = \arg\max_{A} P(A|I, Q)$$

where $P(A|I, Q)$ denotes the probability of (A) being the correct answer given (I) and (Q). Essentially, the aim in each instance is to maximize the probability of obtaining the correct answer based on the provided inputs. This formalism captures the essence of various QA types by emphasizing probabilistic inference from the given data.

Algorithm 3 VQA(I, Q)

Input: Image I, Question Q
Output: Generated answer A

$V \leftarrow$ EncodeImage(I)
$Q\_encoded \leftarrow$ EncodeQuestion(Q)
$R \leftarrow$ Combine(V, $Q\_encoded$)
$A \leftarrow$ empty sequence
while ¬ IsComplete(A) **do**
    $a_i \leftarrow$ GenerateNextToken(R, A)
    Append a_i to A
end while
return A

Overall, VQA encodes the image and question into their respective representations, combines them, and generates the answer token by token based on the combined representation.

As an example, let us envision an image described below, paired with a corresponding question:

I: A photograph showing a cat sitting on a mat.

Q: What is the animal in the picture doing?

Here, The VQA process begins by taking the image and the question as input, employing computer vision techniques to extract relevant features from the image while simultaneously analyzing the question to comprehend its context and intent. Subsequently, the system generates an answer, leveraging both the visual content of the image and the

question posed. For example, it might produce the answer "The animal in the picture is sitting". This instance demonstrates how VQA seamlessly integrates image analysis with NLP to address inquiries about visual content, with potential applications ranging from aiding visually impaired individuals to improving image search engines and enabling interactive assistance in augmented reality contexts

On other words, VQA systems are characterized by several distinctive features. They go beyond simple object recognition by aiming to understand the semantics of both visual content and the questions posed by users, interpreting context, relationships, and attributes within the visual input to generate accurate answers. These systems employ multimodal fusion techniques (Ben-Younes et al., 2017; Fukui et al., 2016; Lao et al., 2021; Gao et al., 2018) to integrate visual and textual information, combining image understanding with NLP. This approach allows VQA systems to achieve a comprehensive grasp of the input data and effectively handle diverse types of questions and visual content (Ben-Younes et al., 2017; Yu et al., 2018).

VQA systems are applied in various domains such as image and scene understanding (Goyal et al., 2017; Biten et al., 2019), assistive technology (Azuma et al., 2022; Biten et al., 2019), and content-based image retrieval (Yu et al., 2020), facilitating tasks like image captioning, visual reasoning, and improving accessibility for the visually impaired. Overall, they represent a significant advancement in AI and NLP, enabling machines to comprehend and respond to questions about visual content in a manner that mimics human-like understanding and paving the way for further research and development.

6.2. Methodologies

The methodologies employed in VQA systems encompass a range of techniques aimed at integrating multiple modalities, such as text and images, to generate answers to questions. Three key methodologies commonly utilized in VQA systems are multimodal fusion, image captioning, and attention mechanisms.

- Multimodal Fusion: Multimodal fusion techniques (Cadene et al., 2019; Ben-Younes et al., 2017; Lu et al., 2023) involve the integration of information from different modalities, such as text and images, to facilitate a deeper understanding of the content. In VQA systems, multimodal fusion enables the combination of textual questions with visual inputs to generate accurate and contextually relevant answers. Various fusion strategies, including early fusion (combining modalities at the input level) and late fusion (integrating modalities at a higher representation level), are employed to effectively merge information from different sources.
- Image Captioning: Image captioning (Anderson et al., 2018) is a technique commonly utilized in VQA systems to generate textual descriptions or captions for images. In the context of QA, image captioning enables the generation of answers based on the visual content depicted in images. VQA systems leverage pre-trained image captioning models, such as those based on CNNs and RNNs, to generate descriptive captions for images, which can then be used to derive answers to questions about the visual content.
- Attention Mechanisms: Attention mechanisms (Vaswani et al., 2017) play a crucial role in VQA systems by enabling the model to focus on relevant regions or features within the input data (Lu et al., 2023). In the context of multimodal question answering, attention mechanisms allow the model to dynamically allocate attention to relevant words in the question and salient features in the visual input. By attending to informative regions or features, attention mechanisms enhance the model's ability to generate accurate and contextually relevant answers. Various attention mechanisms, such as self-attention (Liu et al., 2021) and cross-modal attention (Gong et al., 2021), are employed to capture dependencies between different modalities and facilitate effective information integration.

Additionally, VQA involves a combination of various deep learning architectures tailored to handle both visual and textual data. Here, we will delve into three key techniques commonly used in VQA:

- Convolutional Neural Networks: CNNs have been instrumental in processing visual data, particularly images (Chen et al., 2015b). In the context of VQA, CNNs are employed for extracting visual features from images. These networks consist of convolutional layers that learn hierarchical representations of visual features, capturing information such as edges, textures, and object shapes. Pre-trained CNN models like ResNet, VGG, or Inception are often used as feature extractors in VQA pipelines to encode images into a fixed-dimensional feature representation, which can then be combined with textual features for QA.
- Recurrent Neural Networks: RNNs including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are commonly used for processing sequential data such as textual inputs (Manmadhan and Kovoor, 2020). In VQA, RNNs (Osman and Samek, 2019) are utilized to process and encode textual questions. They capture the sequential dependencies and semantic information present in the questions, enabling the model to understand the context and generate meaningful representations. RNNs can be integrated with CNNs through techniques like attention mechanisms or multimodal fusion to combine visual and textual information effectively for answering questions.
- Graph-Based Models: Graph-based models have gained prominence in VQA tasks, especially in datasets like GQA (Hudson and Manning, 2019; Liang et al., 2021), which require reasoning over structured knowledge graphs. These models represent the relationships between entities (objects, attributes, and relations) in the scene using graph structures. Graph Convolutional Networks (GCNs) or Graph Attention Networks (GATs) are employed to perform reasoning and inference over these graphs. By propagating information across nodes and edges in the graph, these models can infer complex relationships and dependencies, enabling accurate answering of questions that involve spatial, relational, or compositional reasoning.

In summary, techniques for VQA involve leveraging the capabilities of CNNs for visual feature extraction, recurrent networks for processing textual inputs, and graph-based models for reasoning over structured knowledge representations. By integrating these techniques, VQA systems can effectively combine visual and textual information to provide accurate and contextually relevant answers to questions about images. The methodologies for VQA systems, including multimodal fusion (Cadene et al., 2019; Ben-Younes et al., 2017; Lu et al., 2023), image captioning (Anderson et al., 2018), and attention mechanisms (Liu et al., 2021; Gong et al., 2021), enable the integration of textual and visual information to generate accurate and contextually relevant answers to questions across diverse domains. These techniques play a crucial role in advancing the capabilities of QA systems and facilitating effective communication between humans and machines.

6.3. Applications

The dataset for real-world visual reasoning and compositional QA, such as Gqa dataset (Hudson and Manning, 2019), along with all possible associated systems, finds applications in various domains. It primarily leverages its capabilities in image understanding, scene understanding, and assistive technology.

- Image Understanding: VQA enables machines to comprehend images by answering questions related to their content (Ben-Younes et al., 2017; Cadene et al., 2019). This application finds utility in

fields (Wu et al., 2017) like image recognition, where systems can accurately identify objects, people, or scenes depicted in images. For example, a VQA system can answer questions like "What objects are present in the image?" or "How many people are in the picture?" This functionality is vital in image search engines, content moderation systems, and automated image annotation tools.

- Scene Understanding: VQA systems extend their capabilities beyond recognizing individual objects to understanding the context and relationships within a scene (Azuma et al., 2022; Biten et al., 2019). By answering questions about scenes depicted in images, these systems can infer higher-level concepts such as spatial relationships, actions, and interactions between objects. For instance, a VQA system can answer questions like "What is happening in the scene?" or "How are the objects positioned relative to each other?" This capability is valuable in applications such as autonomous driving, robotics, and surveillance systems.
- Assistive Technology: VQA can serve as a valuable tool in assistive technology (Das et al., 2019) for individuals with visual impairments or disabilities. By answering questions about visual content, VQA systems can provide auditory (Li et al., 2022) or textual descriptions of images to assist users in understanding their surroundings (Antol et al., 2015). For example, a visually impaired person can use a VQA-powered mobile app to take a picture and ask questions like "What objects are in front of me?" or "Where is the nearest exit?" This application enhances accessibility and independence for individuals with visual disabilities, empowering them to navigate and interact with the environment more effectively.

In summary, VQA systems have diverse applications in image understanding, scene understanding, and assistive technology, enabling tasks ranging from object recognition and scene interpretation to providing visual assistance for individuals with disabilities. These applications demonstrate the versatility and utility of VQA systems across various domains and highlight their potential to address real-world challenges in visual perception and HCI.

6.4. Evaluation scores and datasets

Evaluation scores and datasets play a crucial role in assessing the performance and benchmarking of QA systems, GQA, VQA (Antol et al., 2015), COCO-QA (Ren et al., 2015), and VizWiz (Gurari et al., 2018; Rocktäschel et al., 2015), more datasets can be found in Sharma and Jalal (2021) and Kafle and Kanan (2017b). These datasets are designed to evaluate the generalization capabilities of QA systems across diverse visual and semantic concepts. It contains images annotated with complex questions that require understanding of both visual content and underlying semantics. more datasets can be found in [78, 173]. Here's an overview of evaluation scores and datasets for each of these QA tasks:

1. Visual Question Answering (VQA):
    - Datasets: VQA datasets typically include images paired with questions and corresponding answers. Common datasets include VQA v1 (Antol et al., 2015) and v2 (Goyal et al., 2017), VizWiz (Gurari et al., 2018; Rocktäschel et al., 2015), and CLEVR (Johnson et al., 2017). VQA v1 and v2 consist of open-ended questions with multiple-choice answers, covering a variety of image categories.
    - Evaluation Scores: For testing purposes as outlined by Antol et al. (2015), two answering modalities are provided: open-answer and multiple-choice. In the open-answer task, responses are evaluated for accuracy based on the agreement of at least three human responses, with adjustments for lowercase formatting, numeric representation, and punctuation removal. Soft scores like Word2Vec are avoided due to their tendency to conflate distinct terms. For the multiple-choice task, 18 answer options are created for each question. The accuracy of chosen options is determined by human response agreement, scaled and clipped similarly to the open-answer task. Candidate answer sets are derived from correct, plausible, popular, and random sources to ensure varied and challenging choices.
2. COCO-QA (Common Objects in COntext QA):
    - Dataset: COCO-QA dataset (Ren et al., 2015) is based on the COCO image dataset and consists of images with associated questions and answers. The questions cover various aspects of the image content, such as object recognition, spatial relationships, and scene understanding.
    - Evaluation Scores: The Wu-Palmer Similarity (WUPS), introduced (Wu and Palmer, 1994), emerged as an alternative evaluation scores to accuracy in the work by Malinowski and Fritz (2014). It aims to gauge the disparity between a predicted answer and the ground truth by assessing their semantic correspondence. WUPS assigns a value between 0 and 1 to indicate their similarity, relying on the concept of the least common subsumer and traversing the semantic tree to measure their closeness. Unlike accuracy, WUPS penalizes semantically similar but not identical words to a lesser extent. For instance, 'bald eagle' and 'eagle' would score 0.96 in similarity, while 'bald eagle' and 'bird' would score 0.88. Nonetheless, WUPS sometimes assigns relatively high scores even to distant concepts, as seen in the example of 'raven' and 'writing desk' with a score of 0.4. To address this issue, Malinowski and Fritz (2014) proposed a modification to WUPS, suggesting a thresholding mechanism. Scores below a specified threshold are adjusted by a scaling factor. They recommended a threshold of 0.9 and a scaling factor of 0.1. This refined version of WUPS has become the standard measure for assessing performance on datasets like DAQUAR (Malinowski and Fritz, 2014) and COCO-QA, alongside simple accuracy.
3. VizWiz:
    - Dataset: VizWiz dataset focuses on VQA for visually impaired users. It comprises images captured by blind individuals along with questions related to the image content. The questions often require reasoning about everyday scenes and objects.
    - Evaluation Scores: The VizWiz dataset utilizes four evaluation scores: Accuracy (Antol et al., 2015), CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). While Accuracy is particularly useful for single-word answers, it falls short when dealing with the dataset's abundance of multi-word responses. In such instances, the image description scores introduced by Chen et al. (2015a), tailored for assessing longer phrases or sentences, are deemed more suitable.

In summary, evaluation scores and datasets for VQA, COCO-QA, and VizWiz are essential for assessing the performance and generalization capabilities of QA systems across various domains and modalities. These datasets provide valuable resources for benchmarking QA models and advancing research in the field of question answering.

6.5. Challenges and limitations

The VQA task presents several challenges and limitations that impact its performance and effectiveness. These challenges primarily revolve around semantic understanding, commonsense reasoning, and dataset biases:

- Semantic Understanding: VQA systems often struggle with understanding the nuanced semantics of natural language questions (Shah et al., 2019a). While they may excel at basic keyword matching and syntactic analysis, deeper semantic understanding remains a significant challenge. This limitation can lead to inaccuracies and misunderstandings, particularly when questions involve complex or ambiguous language. For example, if the question is "What is the man holding?" the system might fail to identify the object when the person in the image is holding up their goggles. It may also make incorrect guesses based on the surrounding context, which would need to be corrected through further explanation. This issue occurs because the question may be vague, there may not be enough context, or the system might struggle to grasp subtle details from other parts of the image.
- Commonsense Reasoning: Many questions in the VQA task require reasoning beyond explicit information provided in the text, especially those questions that require multi-step reasoning or an understanding of spatial relationships. They necessitate the application of commonsense knowledge and reasoning abilities (Ravi et al., 2023) to derive correct answers. VQA systems often lack robust commonsense reasoning capabilities, which can result in incorrect or nonsensical responses to questions that require implicit knowledge or logical inference. This issue arises because many VQA models lack advanced reasoning capabilities and tend to focus too much on basic image features, causing them to overlook the more complex relationships within the image.
- Dataset Biases: VQA systems encounter several challenges, including a lack of world knowledge and common sense, which leads to incorrect answers when external understanding is required. Biases in training data can result in biased responses, such as incorrectly assuming a woman is in the kitchen based on skewed training images. VQA datasets are particularly vulnerable to biases introduced during data collection or annotation, which can appear as imbalanced question distributions, over-representation of certain answer types, or unclear annotations (Manjunatha et al., 2019). These biases can negatively impact the performance of VQA systems, causing them to favor biased solutions instead of demonstrating true understanding or reasoning abilities. Additionally, VQA systems often struggle with fine-grained visual understanding, leading to misidentifications of similar objects, such as apples and oranges. Furthermore, over-reliance on image-text matching can cause errors, like associating the color of a dog with the dominant background color instead of the dog itself.

Addressing these challenges and limitations requires advancements in NLU, commonsense reasoning, and dataset curation techniques. Researchers are actively exploring novel approaches, such as incorporating external knowledge sources, developing more sophisticated reasoning models, and implementing bias mitigation strategies, to improve the performance and robustness of VQA systems. Additionally, ongoing efforts to create more diverse and balanced datasets can help mitigate biases and promote the development of more reliable and inclusive VQA models.

6.6. Recent advancements and research trends

Recent advancements and research trends with VQA have seen significant progress in enhancing the performance and capabilities of VQA systems. GQA is a large-scale dataset designed to evaluate the reasoning abilities of VQA models by posing questions about images. Here are some key recent advancements and research trends with VQA:

- Multimodal Reasoning: One of the primary research trends with VQA is advancing multimodal reasoning capabilities. Researchers (Wang et al., 2023; Sima et al., 2024; Wang et al., 2020b; Nourali and Dolkhani, 2024) are exploring techniques to enable models to reason jointly across different modalities, such as images and text, to answer complex questions that require understanding both visual and textual information, especially those that involve multi-step reasoning and question decomposition.
- Semantic Understanding: Recent research has focused on improving the semantic understanding (Shah et al., 2019a) of questions and images within the GQA dataset. This involves developing models capable of capturing the nuanced relationships between objects, attributes, and spatial configurations within images, enabling more accurate and contextually relevant answers.
- Attention Mechanisms: Attention-based mechanisms, such as visual-linguistic attention or dual self-attention (Liu et al., 2021; Gong et al., 2021) play a crucial role in VQA systems by allowing models to focus on relevant regions of an image or words in a question when generating answers. Recent advancements have explored novel attention mechanisms tailored to the specific challenges posed by the VQA dataset, leading to improved performance and interpretability.
- Transfer Learning: Transfer learning techniques (Abacha et al., 2019) have gained prominence in VQA research, leveraging pre-trained models on large-scale image and language datasets to bootstrap performance on the GQA task. Fine-tuning strategies and architectural adaptations have been explored to adapt pre-trained models to the nuances of the GQA dataset effectively. Additionally, incorporating external knowledge — such as knowledge graphs, pre-trained language models, or common sense reasoning modules — enables the VQA system to answer questions that require knowledge beyond the training data (Gardères et al., 2020).
- Explainability and Interpretability: With the increasing complexity of VQA models, there is growing interest in enhancing the explainability (Li et al., 2018) and interpretability (Cao et al., 2019) of model predictions on the VQA task. Recent research has focused on developing techniques to generate human-understandable explanations for model decisions, enabling users to understand how and why a particular answer was generated.
- Dataset Expansion and Benchmarking: Continuous efforts are underway to expand the VQA dataset and introduce new benchmarks that reflect real-world challenges in VQA (Masry et al., 2022; Wu et al., 2017). These efforts include creating more diverse, balanced, and representative datasets, applying data augmentation techniques, developing complex question-answer pairs, and introducing adversarial examples to evaluate model robustness.

Overall, recent advancements and research trends with VQA have propelled the development of more sophisticated and capable VQA systems, paving the way for applications in areas such as image understanding, scene understanding, and HCI.

7. Comparison of QA paradigms

EQA, GQA, and VQA paradigms represent different approaches to answering questions, each with distinct methodologies, and applications as summarized in Table 1.

EQA systems rely on retrieving answers directly from a given corpus of text or data through keyword matching and document retrieval techniques. These systems use methods such as keyword matching, document retrieval, and pattern recognition to identify and extract relevant information from textual data. EQA systems are suitable for tasks like IR, fact-checking, document summarization, and search engines, where the goal is to extract specific information from text sources.

GQA systems, on the other hand, generate answers to questions using NLG techniques. They often synthesize information from various sources or understand and reason about the question context. GQA systems employ natural language understanding, generation, and

Table 1
Comparison of EQA, GQA, and VQA Systems - Methodologies and applications.

System type	Methodologies	Applications
EQA	Uses keyword matching, document retrieval, and pattern recognition to extract answers directly from a text corpus. These methods focus on identifying and retrieving specific information based on exact or pattern-based matches within the text.	Well-suited for applications such as IR, factual-question an, document summarization, and search engines, where precise information extraction from text is essential.
GQA	Leverages NLU, generation, and reasoning to synthesize answers from multiple sources or interpret the context of the question. Techniques include understanding the question, generating relevant responses, and ensuring coherence and contextual accuracy.	Primarily used in chatbots, virtual assistants, dialogue systems, and customer service platforms, where producing contextually appropriate, human-like responses is crucial.
VQA	Combines computer vision and natural language processing to answer questions about images or visual content. This involves image processing, feature extraction, deep learning, and multimodal fusion to interpret visual data and generate textual answers based on both visual and textual inputs.	Applied in fields like image captioning, content-based image retrieval, visual assistance for the visually impaired, autonomous vehicles, and medical image analysis, enabling users to query visual content and receive relevant, context-based responses.

reasoning techniques to comprehend the question context, generate answers, and ensure coherence and relevance. These systems are commonly used in applications such as chatbots, virtual assistants, dialogue systems, and customer service platforms, where natural language understanding and generation are essential for providing human-like responses.

VQA systems answer questions about images or visual content by combining computer vision techniques with NLP to understand both the visual input and the textual question. VQA systems utilize image processing, feature extraction, deep learning, and multimodal fusion techniques to analyze visual content, understand the question, and generate appropriate textual answers. They are used in areas like image captioning, content-based image retrieval, visual assistance for the visually impaired, autonomous vehicles, and medical image analysis, enabling users to ask questions about visual content and receive relevant answers.

In summary, EQA, GQA, and VQA paradigms offer different approaches to question answering, each tailored to specific domains and applications. While EQA systems excel at extracting information from text, GQA systems focus on generating coherent responses, and VQA systems handle questions about visual content by integrating image understanding with NLP. Understanding the methodologies, and applications of these paradigms is crucial for designing effective question-answering systems for various tasks and domains.

7.1. The strengths and weaknesses of each paradigm

Extractive Question Answering (EQA) systems have several strengths and weaknesses. Among their strengths, EQA systems are often computationally efficient since they rely on simple keyword-matching techniques to extract answers directly from the text corpus. This makes them transparent, as the extracted answers come directly from the source text, providing traceability in the IR process. Additionally, EQA systems are well-suited for answering fact-based questions where the answer is explicitly stated in the text corpus. However, EQA systems also have weaknesses. They have a limited understanding of the context or semantics of the text, which can result in inaccuracies when answering questions that require deeper comprehension. They struggle with answering questions that are ambiguous or require reasoning beyond simple keyword matching. Furthermore, the accuracy of EQA systems heavily depends on the quality and relevance of the text corpus, with inaccuracies or biases in the source text potentially leading to erroneous answers.

Generative Question Answering (GQA) systems also come with their own set of strengths and weaknesses. In terms of strengths, GQA systems have the ability to understand context and semantics, allowing them to generate answers that go beyond simple extraction. They are flexible and capable of generating answers to a wide range of questions, including those that require reasoning, inference, or synthesis of information from multiple sources. GQA systems are also better equipped to handle ambiguous questions by leveraging contextual cues and linguistic knowledge. On the downside, building and training GQA systems can be complex and resource-intensive due to the need for advanced NLU and generation techniques. They often require large amounts of annotated training data to learn the complex patterns and relationships in natural language. Additionally, the generated answers may lack transparency, making it challenging to understand how the system arrived at a particular response.

VQA systems integrate visual information from images or videos with textual queries, enabling them to answer questions that involve both visual and textual context. This multimodal understanding is one of their strengths, allowing them to provide richer and more informative answers by leveraging visual cues and context in addition to textual information. VQA systems have applications in image understanding, scene understanding, and assistive technologies, making them versatile tools for various domains. However, VQA systems also face challenges. Processing and understanding visual information is inherently more complex than text, requiring sophisticated computer vision techniques and models. Integrating visual and textual information effectively poses challenges in terms of data representation, fusion techniques, and model architecture design. Moreover, VQA systems often require large-scale datasets with paired visual and textual information for training, which may be limited or expensive to obtain.

Overall, each paradigm has its pros and cons as shown in Table 2, and the choice of approach depends on the specific requirements and constraints of the task at hand.

7.2. Trade-offs and considerations for selecting the appropriate QA approach

When selecting the appropriate QA approach, various trade-offs and considerations should be taken into account. These include task requirements, accuracy and efficiency, data availability and quality, interpretability, scalability, domain and language specificity, as well as ethical and bias considerations, as delineated in Table 3.

Understanding the specific requirements of a QA task is crucial. It is important to consider whether the task involves answering fact-based questions, generating novel responses, handling multimedia inputs, or understanding complex contexts. Different QA approaches may be more suitable for different types of tasks.

Moreover, there are trade-offs between accuracy and efficiency in QA systems. GQA may offer more accurate and contextually relevant answers, but they often require more computational resources and time for training and inference. Conversely, EQA may sacrifice some accuracy for efficiency, providing quick and straightforward answers.

Furthermore, the availability and quality of training data play a significant role in selecting the appropriate QA approach. GQA typically require large amounts of high-quality training data to learn to generate accurate responses, while EQA may be more robust to data scarcity and noise, relying on matching existing information.

Considerably, the interpretability of the QA system's outputs is crucial, especially in critical applications where understanding the reasoning behind the answers is essential. GQA may offer more transparent

Table 2
The pros and cons of QA paradigms.

QA Type	Pros	Cons
EQA	**Efficiency**: Efficient in terms of computation due to its reliance on basic keyword matching techniques. **Transparency**: Provides answers directly sourced from the text. **Suitability for factual questions**: Ideal for answering straightforward factual questions where answers are clearly stated in the text corpus.	**Limited Understanding**: Lacks deep contextual comprehension or semantic reasoning. **Struggles with ambiguity**: Has difficulty handling ambiguous questions or those requiring higher-level reasoning beyond keyword matching. **Text Quality Dependence**: Accuracy is highly dependent on the quality and relevance of the input text corpus.
GQA	**Contextual comprehension**: Capable of understanding context and semantics in a more nuanced way. **Flexibility**: Can handle a broad range of queries, including those needing reasoning, inference, or synthesis from multiple sources. **Ambiguity resolution**: More adept at addressing ambiguous questions by utilizing contextual and linguistic cues.	**Complexity**: Developing and training GQA systems is resource-intensive and complex. **Data needs**: Requires large annotated datasets for training. **Lack of interpretability**: Generated responses may lack transparency, making it difficult to understand the reasoning behind the system's answers.
VQA	**Multimodal capability**: Can process both visual and textual data simultaneously. **Enhanced Understanding**: Provides richer, more detailed answers by incorporating visual context alongside text. **Applications in visual contexts**: Particularly effective for tasks involving image interpretation, scene understanding, and assistive technologies.	**Challenges in visual understanding**: Requires advanced computer vision techniques for accurate interpretation of visual data. **Integration issues**: Effectively merging visual and textual data presents significant challenges. **Data constraints**: Needs large, high-quality datasets with paired visual and textual information, which may be scarce or costly to obtain.

Table 3
Key considerations and trade-offs in selecting the appropriate QA approach.

Aspect	Explanation
Task requirements	It is crucial to understand the specific requirements of the QA task. This involves determining whether the task requires answering factual questions, generating novel responses, handling multimodal inputs, or understanding complex contexts. The selection of the QA approach should be aligned with the task type, as different methods excel in different contexts.
Accuracy vs. Efficiency	There is an inherent trade-off between accuracy and efficiency in QA systems. GQA often produce more precise and contextually relevant answers, but they typically require more computational resources and time for both training and inference. EQA, on the other hand, prioritize efficiency, providing quicker, more straightforward answers at the cost of some accuracy.
Data availability and Quality	The performance of QA approaches is strongly influenced by the quality and availability of training data. GQA, which create answers from scratch, require large, high-quality datasets to perform effectively. EQA can be more robust in data-scarce or noisy environments, as they rely on selecting relevant information from existing text rather than generating new content.
Interpretability	The ability to interpret answers produced by QA systems is especially important in critical contexts. GQA, which directly produce responses, can offer clearer outputs, while EQA, relying on ranking or matching algorithms, may be less transparent.
Scalability	Scalability is a critical factor, particularly in systems that handle large volumes of data or respond to real-time queries. It is crucial to assess if the chosen QA approach can adapt efficiently to expanding data sizes and rising user demands while maintaining optimal performance.
Domain and language specificity	Different QA approaches may excel in particular domains or languages due to variations in vocabulary, syntax, or semantics. Assess the domain knowledge necessary for training and refining QA models, and evaluate if the chosen approach can generalize effectively across diverse domains or languages.
Ethical and bias considerations	Ethical considerations, including fairness, bias, and transparency, should be central in selecting a QA approach. It is important to examine potential biases in the training data and model responses, ensuring that the chosen method minimizes bias and promotes fairness and inclusivity in its outputs.

outputs by explicitly generating responses, whereas EQA may provide answers based on opaque ranking or matching algorithms.

Additionally, scalability is another important consideration, especially for applications processing large data volumes or handling real-time queries. It is crucial to evaluate whether the selected QA approach can effectively scale to accommodate growing data sizes and increasing user demands without compromising performance.

Moreover, certain QA approaches may perform better in specific domains or languages due to differences in vocabulary, syntax, or semantics. It is important to consider the domain expertise required for training and fine-tuning QA models and whether the selected approach can generalize well across different domains or languages.

Lastly, ethical considerations such as fairness, bias, and transparency should be evaluated in the QA process. It is essential to assess potential biases in the training data and model outputs and ensure that the selected approach mitigates bias and promotes fairness and inclusivity

By carefully weighing these trade-offs and considerations, stakeholders can select the most appropriate QA approach that aligns with their specific requirements, constraints, and ethical considerations.

8. Challenges and future directions

8.1. Common challenges across EQA, GQA, and VQA

While EQA, GQA, and VQA each target different domains, they share several fundamental challenges related to NLU, multi-modal data integration, and the generation of reliable answers. Below are key challenges that span these paradigms:

1. Semantic Understanding and Contextual Awareness: Understanding the semantic meaning of questions, along with the context in which they arise, is a critical challenge. For example, even slight differences in phrasing or ambiguous terms can lead to significantly different interpretations of a question. This challenge becomes more complex when questions are posed across different modalities (text, images, or videos). The ability to discern context and intent from complex or imprecise language remains a foundational hurdle.
2. Handling Ambiguity and Variability: Natural language is inherently ambiguous, and this poses a problem for QA systems, particularly when the input data is noisy or unstructured. Ambiguity

may be lexical (same word having different meanings), syntactic (ambiguous sentence structure), or semantic (interpretation depending on context). Systems must be robust enough to disambiguate these variations to generate accurate answers. Moreover, different languages, dialects, or cultural contexts further exacerbate these challenges.

3. Cross-Modal Integration: Particularly for VQA and GQA, the challenge of integrating visual data with textual input is pronounced. In VQA, for instance, answering a question about an image requires bridging the gap between visual and textual modalities. Extracting relevant visual features (e.g., objects, scenes, spatial relations) and aligning them with the textual content requires novel methods for joint multimodal representations. This issue also extends to GQA, where generalization across different domains may demand more advanced techniques for aligning knowledge graphs or structured information with unstructured data sources.
4. Scalability and Computational Efficiency: As datasets grow larger and more diverse, scalability becomes an essential concern. Both EQA and GQA systems must be capable of processing vast amounts of data quickly while maintaining accuracy. Efficient indexing, retrieval, and inference algorithms are necessary to ensure that performance does not degrade as the scope of questions or the scale of the knowledge base increases.
5. Robustness to Errors and Noisy Inputs: Many QA systems encounter noisy, incomplete, or erroneous inputs. For example, in EQA, questions may contain spelling mistakes or poorly structured sentences. In VQA, images may contain occluded objects or unusual scenes. QA systems must be resilient to such noise while providing answers that are both accurate and reasonable. Ensuring systems remain reliable in real-world, dynamic environments is a challenge that extends across all QA paradigms.
6. Hallucination in Responses: Hallucination, or the generation of incorrect or fabricated answers (Huang et al., 2023; Zhang et al., 2023), is particularly problematic for LLM-driven systems, which are often used in EQA, GQA, and VQA tasks. When LLMs or other neural models infer answers without proper grounding or reasoning over the available data, they can produce misleading results. This issue is critical for maintaining the credibility and trustworthiness of QA systems in applications where factual accuracy is paramount.
7. Evaluation Metrics and Performance Assessment: Assessing the performance of QA systems goes beyond traditional metrics like accuracy or F1 score. In complex systems like GQA or VQA, where answers may vary in complexity, creativity, or abstraction, new evaluation metrics are needed. These should account for system understanding of nuanced questions, reasoning abilities, and the quality of the answer in the given context, rather than merely evaluating whether a "correct" answer is given.

Addressing these challenges requires advancements in NLP, computer vision, knowledge representation, and ML techniques. Interdisciplinary research efforts focusing on these areas can lead to more robust and effective question-answering systems across different modalities.

8.2. Exploring emerging trends, opportunities, and open questions in QA research

Emerging trends and opportunities in QA research are continually shaping the field, offering novel avenues for exploration and innovation. Moreover, the open research questions and areas for future exploration in the field of QA systems encompass a wide range of topics aimed at advancing these systems' capabilities, effectiveness, and applicability. Some key areas for notable trends and opportunities and further investigation include:

- Improving Contextual and Semantic Understanding: Improving semantic understanding in both structured and unstructured contexts is critical. Future research can delve into advanced pretrained language models that better capture the nuances of contextual information. Neuro-symbolic reasoning, which combines symbolic logic with neural networks, could be a key technique for ensuring deeper, more interpretable understanding of questions and answers, particularly in structured domains such as GQA.
- Enhanced Multimodal Fusion and Joint Learning: Multimodal QA systems need to better integrate visual, textual, and potentially auditory data in a more unified manner. Advanced deep learning techniques like cross-attention mechanisms, transformers, and graph neural networks offer promising pathways for fusing different modalities and learning joint representations that enhance understanding. Future research could explore techniques for unsupervised or few-shot learning across multiple modalities to enable better generalization in VQA and GQA settings.
- Explainability and Transparency: As AI-based systems take on more decision-making roles, it becomes essential for them to offer transparent and explainable outputs. Research into explainable AI (XAI) techniques will be crucial to make EQA, GQA, and VQA systems more understandable to end-users, allowing them to trust and interpret the reasoning behind the answers provided. Explainability will be particularly significant in safety-critical systems where understanding the model's reasoning process can help identify and mitigate errors.
- Domain-Specific Adaptation and Knowledge Customization: Developing domain-specific QA systems (e.g., medical, legal, or scientific domains) that can leverage expert knowledge more effectively will become increasingly important. Transfer learning and domain adaptation techniques could enable QA systems to more accurately address questions specific to a given field. This would also require the creation of domain-specific datasets and evaluation benchmarks to test QA systems' performance in these settings.
- Fairness, Bias Mitigation, and Ethical QA: As QA systems become more integrated into real-world applications, ensuring fairness and mitigating biases will be crucial. Research could focus on developing frameworks for detecting and correcting bias in data and models, ensuring that EQA, GQA, and VQA systems generate answers that are equitable and non-discriminatory. Ethical considerations, such as the system's ability to respect privacy and avoid harmful stereotypes, should also be a focal point for future development.
- Interactive and Iterative QA Systems: Moving beyond one-shot question answering, the next generation of QA systems should aim for interactive dialogue systems that can ask clarification questions and refine answers based on user feedback. By fostering an iterative exchange, systems can overcome ambiguities or errors that arise from initial queries and provide more precise, contextually appropriate answers.
- Reducing Hallucination and Enhancing Verifiability: To combat hallucination, future research must explore better grounding mechanisms that tie responses to verifiable sources. Fact-checking models or retrieval-augmented QA systems could play an essential role here. Ensuring that answers are based on reliable, verified data sources will significantly improve system trustworthiness, particularly in high-stakes environments such as healthcare or legal settings.

Overall, these emerging trends and opportunities in QA research reflect the field's dynamism and its ongoing efforts to advance the state-of-the-art in NLP and IR. By embracing these trends and addressing associated challenges, researchers can drive innovation and create more robust, reliable, and responsible QA systems for diverse applications and users. In other words, addressing these open research questions

and exploring new frontiers in QA systems will contribute to advancing the field's capabilities and realizing its potential to address real-world challenges across various domains and applications.

9. Conclusion

This survey offers an in-depth examination of question-answering systems, detailing their evolution and key methodologies, evaluation techniques, characteristics, applications, strengths, limitations, and challenges. It also explores future research directions. By integrating existing research and literature, we provide valuable insights into the core principles and techniques of various paradigms, including extractive, generative, and VQA.

Our discussion highlights the critical role of interdisciplinary collaboration among NLP, computer vision, and ML in advancing question-answering technologies. By pinpointing gaps in current knowledge and suggesting areas for further research, this paper acts as a guide for future exploration in the field. Our findings offer clarity on the capabilities and limitations of different question-answering approaches, aiding in informed decision-making for selecting or designing systems tailored to specific tasks or applications.

In essence, this survey paper seeks to enhance understanding, drive innovation, and advance knowledge in question-answering systems. By presenting recent methodological advancements and identifying ongoing challenges, we aim to provide a valuable reference for researchers and practitioners alike.

CRediT authorship contribution statement

Amer Farea: Writing – review & editing, Writing – original draft, Conceptualization. **Frank Emmert-Streib:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. VQA-Med: Overview of the medical visual question answering task at imageclef 2019. CLEF (Work. Notes) 2 (6).

Abacha, A.B., Zweigenbaum, P., 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Inf. Process. Manage. 51 (5), 570–594.

Abbasiantaeb, Z., Momtazi, S., 2021. Text-based question answering from information retrieval and deep neural network perspectives: A survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 11 (6), e1412.

Abdel-Nabi, H., Awajan, A., Ali, M.Z., 2023. Deep learning-based question answering: a survey. Knowl. Inf. Syst. 65 (4), 1399–1485.

Ahmed, M., Khan, H.U., Iqbal, S., Althebyan, Q., 2022. Automated question answering based on improved tf-idf and cosine similarity. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security. SNAMS, IEEE, pp. 1–6.

Al Mahmud, T., Hussain, M.G., Kabir, S., Ahmad, H., Sobhan, M., 2020. A keyword based technique to evaluate broad question answer script. In: Proceedings of the 2020 9th International Conference on Software and Computer Applications. pp. 167–171.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. 35, 23716–23736.

Allam, A.M.N., Haggag, M.H., 2012. The question answering systems: A survey. Int. J. Res. Rev. Inf. Sciences (IJRRIS) 2 (3).

Alruqi, T.N., Alzahrani, S.M., 2023. Evaluation of an arabic chatbot based on extractive question-answering transfer learning and language transformers. AI 4 (3), 667–691.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086.

Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433.

Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., Sidorov, G., 2019. Unsupervised sentence representations as word information series: Revisiting TF–IDF. Comput. Speech Lang. 56, 107–129.

Assem, H., Sarkar, R., Dutta, S., 2021. QASAR: self-supervised learning framework for extractive question answering. In: 2021 IEEE International Conference on Big Data (Big Data). IEEE, pp. 1797–1808.

Atanasova, P., 2024. Generating fact checking explanations. In: Accountable and Explainable Methods for Complex Reasoning over Text. Springer, pp. 83–103.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., et al., 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. Nat. Mach. Intell. 6 (8), 852–863.

Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M., 2022. Scanqa: 3d question answering for spatial scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19129–19139.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al., 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Bai, Y., Wang, D.Z., 2021. More than reading comprehension: A survey on datasets and metrics of textual question answering. arXiv preprint arXiv:2109.12264.

Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/Or Summarization. pp. 65–72.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al., 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

Bates, M., 1995. Models of natural language understanding. Proc. Natl. Acad. Sci. 92 (22), 9977–9982.

Ben-Younes, H., Cadene, R., Cord, M., Thome, N., 2017. Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2612–2620.

Berant, J., Chou, A., Frostig, R., Liang, P., 2013. Semantic parsing on freebase from question-answer pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1533–1544.

Bhaskar, P., Banerjee, S., Pakray, P., Banerjee, S., Bandyopadhyay, S., Gelbukh, A., 2013. A hybrid question answering system for Multiple Choice Question (MCQ). In: Question Answering for Machine Reading Evaluation (QA4MRE) At CLEF 2013 Conference and Labs of the Evaluation Forum.

Bidgoly, A.J., Amirkhani, H., Baradaran, R., 2022. Clustering-based sequence to sequence model for generative question answering in a low-resource language. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 22 (2), 1–14.

Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D., 2019. Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4291–4301.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Butt, S., Ashraf, N., Siddiqui, M.H.F., Sidorov, G., Gelbukh, A., 2021. Transformer-based extractive social media question answering on TweetQA. Comput. Sist. 25 (1), 23–32.

Cadene, R., Ben-Younes, H., Cord, M., Thome, N., 2019. Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1989–1998.

Cao, Q., Liang, X., Li, B., Lin, L., 2019. Interpretable visual question answering by reasoning on dependency trees. IEEE Trans. Pattern Anal. Mach. Intell. 43 (3), 887–901.

Cao, Q., Trivedi, H., Balasubramanian, A., Balasubramanian, N., 2020. DeFormer: Decomposing pre-trained transformers for faster question answering. arXiv preprint arXiv:2005.00697.

Castel, O., Ram, O., Efrat, A., Levy, O., 2021. How optimal is greedy decoding for extractive question answering?. arXiv preprint arXiv:2108.05857.

Chang, Y.-H., Guo, Y.-T., Fu, L.-C., Chiu, M.-J., Chiu, H.-M., Lin, H.-J., 2023. Interactive healthcare robot using attention-based question-answer retrieval and medical entity extraction models. IEEE J. Biomed. Heal. Inform..

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L., 2015a. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M., 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18030–18040.

Chen, A., Stanovsky, G., Singh, S., Gardner, M., 2019. Evaluating question answering evaluation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 119–124.

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., Nevatia, R., 2015b. Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960.

Chen, Y., Zulkernine, F., 2021. BIRD-QA: a BERT-based information retrieval approach to domain specific question answering. In: 2021 IEEE International Conference on Big Data (Big Data). IEEE, pp. 3503–3510.

Cheng, H., Shen, Y., Liu, X., He, P., Chen, W., Gao, J., 2021. UnitedQA: A hybrid approach for open domain question answering. arXiv preprint arXiv:2101.00178.

Chheang, V., Sharmin, S., Márquez-Hernández, R., Patel, M., Rajasekaran, D., Caulfield, G., Kiafar, B., Li, J., Kullu, P., Barmaki, R.L., 2024. Towards anatomy education with generative AI-based virtual assistants in immersive virtual reality environments. In: 2024 IEEE International Conference on Artificial Intelligence and EXtended and Virtual Reality (AIxVR). IEEE, pp. 21–30.

Chiche, A., Yitagesu, B., 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. J. Big Data 9 (1), 10.

Chiticariu, L., Li, Y., Reiss, F., 2013. Rule-based information extraction is dead! long live rule-based information extraction systems!. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 827–832.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2023. Palm: Scaling language modeling with pathways. J. Mach. Learn. Res. 24 (240), 1–113.

Chung, Y.-A., Lee, H.-Y., Glass, J., 2017. Supervised and unsupervised transfer learning for question answering. arXiv preprint arXiv:1711.05345.

Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., Palomaki, J., 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. Trans. Assoc. Comput. Linguist. 8, 454–470.

Cohen, P.R., Perrault, C.R., Allen, J.F., 2014. Beyond question answering. In: Strategies for Natural Language Processing. Psychology Press, pp. 245–274.

Cortes, E.G., Woloszyn, V., Barone, D.A., 2018. When, where, who, what or why? a hybrid model to question answering systems. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. Springer, pp. 136–146.

Cui, H., Sun, R., Li, K., Kan, M.-Y., Chua, T.-S., 2005. Question answering passage retrieval using dependency relations. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 400–407.

Cui, W., Xiao, Y., Wang, H., Song, Y., Hwang, S.-w., Wang, W., 2019. KBQA: learning question answering over QA corpora and knowledge bases. arXiv preprint arXiv:1903.02419.

Das, A., Anjum, S., Gurari, D., 2019. Dataset bias: A case study for visual question answering. Proc. Assoc. Inf. Sci. Technol. 56 (1), 58–67.

Das, R., Zaheer, M., Reddy, S., McCallum, A., 2017. Question answering on knowledge bases and text using universal schema and memory networks. arXiv preprint arXiv:1704.08384.

de Winter, J.C., 2023. Can ChatGPT pass high school exams on English Language Comprehension? Int. J. Artif. Intell. Educ. 1–16.

Demner-Fushman, D., Lin, J., 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. pp. 841–848.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dhingra, B., Mazaitis, K., Cohen, W.W., 2017. Quasar: Datasets for question answering by search and reading. arXiv preprint arXiv:1707.03904.

Diefenbach, D., Lopez, V., Singh, K., Maret, P., 2018. Core techniques of question answering systems over knowledge bases: a survey. Knowl. Inf. Syst. 55, 529–569.

Er, N.P., Cicekli, I., 2013. A factoid question answering system using answer pattern matching. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing. pp. 854–858.

Fader, A., Zettlemoyer, L., Etzioni, O., 2014. Open question answering over curated and extracted knowledge bases. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1156–1165.

Farea, A., Emmert-Streib, F., 2024. Experimental design of extractive question-answering systems: Influence of error scores and answer length. J. Artificial Intelligence Res. 80, 87–125.

Farea, A., Yang, Z., Duong, K., Perera, N., Emmert-Streib, F., 2022. Evaluation of question answering systems: Complexity of judging a natural language. arXiv preprint arXiv:2209.12617.

Fares, M., Oepen, S., Zhang, Y., 2013. Machine learning for high-quality tokenization replicating variable tokenization schemes. In: Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I 14. Springer, pp. 231–244.

Feng, Y., Chen, X., Lin, B.Y., Wang, P., Yan, J., Ren, X., 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. arXiv preprint arXiv:2005.00646.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al., 2010. Building Watson: An overview of the DeepQA project. AI Mag. 31 (3), 59–79.

Floridi, L., Chiriatti, M., 2020. GPT-3: Its nature, scope, limits, and consequences. Minds Mach. 30, 681–694.

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

Gaizauskas, R.J., Humphreys, K., 2000. A combined IR/NLP approach to question answering against large text collections.. In: RIAO. pp. 1288–1304.

Galitsky, B., 2025. Truth-o-meter: Collaborating with llm in fighting its hallucinations. In: Interdependent Human-Machine Teams. Elsevier, pp. 175–210.

Gao, P., Li, H., Li, S., Lu, P., Li, Y., Hoi, S.C., Wang, X., 2018. Question-guided hybrid convolution for visual question answering. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 469–485.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. 2, arXiv preprint arXiv:2312.10997.

Gardères, F., Ziaeefard, M., Abeloos, B., Lecue, F., 2020. Conceptbert: Concept-aware representation for visual question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 489–498.

Gong, H., Chen, G., Liu, S., Yu, Y., Li, G., 2021. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 456–460.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913.

Gunasekara, C., Feigenblat, G., Sznajder, B., Aharonov, R., Joshi, S., 2021. Using question answering rewards to improve abstractive summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 518–526.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR, pp. 1321–1330.

Guo, M., Zhang, M., Reddy, S., Alikhani, M., 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In: 3rd Conference on Automated Knowledge Base Construction.

Gupta, P., Gupta, V., 2012. A survey of text question answering techniques. Int. J. Comput. Appl. 53 (4).

Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P., 2018. Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617.

Hardalov, M., Koychev, I., Nakov, P., 2018. Towards automated customer support. In: Artificial Intelligence: Methodology, Systems, and Applications: 18th International Conference, AIMSA 2018, Varna, Bulgaria, September 12–14, 2018, Proceedings 18. Springer, pp. 48–59.

Hirschman, L., Gaizauskas, R., 2001. Natural language question answering: the view from here. Nat. Lang. Eng. 7 (4), 275–300.

Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., Abend, O., 2021. Q˙ {2} : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. arXiv preprint arXiv:2104.08202.

Hovy, E.H., Gerber, L., Hermjakob, U., Junk, M., Lin, C.-Y., 2000. Question answering in webclopedia. In: TREC. Vol. 52, pp. 53–56.

Hu, Z., Yang, P., Li, B., Sun, Y., Yang, B., 2023. Biomedical extractive question answering based on dynamic routing and answer voting. Inf. Process. Manage. 60 (4), 103367.

Hua, S., Jin, S., Jiang, S., 2024. The limitations and ethical considerations of ChatGPT. Data Intell. 6 (1), 201–239.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al., 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232.

Hudson, D.A., Manning, C.D., 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6700–6709.

Imran, M., Almusharraf, N., 2024. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. Smart Learn. Environ. 11 (1), 22.

Iyer, S., Min, S., Mehdad, Y., Yih, W.-t., 2021. RECONSIDER: improved re-ranking using span-focused cross-attention for open domain question answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1280–1287.

Izacard, G., Grave, E., 2020. Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P., 2023. Survey of hallucination in natural language generation. ACM Comput. Surv. 55 (12), 1–38.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., Yu, S., 2022. Biomedical question answering: a survey of approaches and challenges. ACM Comput. Surv. 55 (2), 1–36.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R., 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910.

Kafle, K., Kanan, C., 2017a. An analysis of visual question answering algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1965–1973.

Kafle, K., Kanan, C., 2017b. Visual question answering: Datasets, algorithms, and future challenges. Comput. Vis. Image Underst. 163, 3–20.

Kamsties, E., Peach, B., 2000. Taming ambiguity in natural language requirements. In: Proceedings of the Thirteenth International Conference on Software and Systems Engineering and Applications. Vol. 1315.

Keyvan, K., Huang, J.X., 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. ACM Comput. Surv. 55 (6), 1–40.

Kim, M., Lee, S., Oh, Y., Choi, H., Kim, W., 2020. A near-real-time answer discovery for open-domain with unanswerable questions from the web. IEEE Access 8, 158346–158355.

Klakow, D., Peters, J., 2002. Testing the correlation of word error rate and perplexity. Speech Commun. 38 (1–2), 19–28.

Klein, T., Nabi, M., 2019. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. arXiv preprint arXiv:1911.02365.

Kobayashi, M., Takeda, K., 2000. Information retrieval on the web. ACM Comput. Surv. 32 (2), 144–173.

Kolomiyets, O., Moens, M.-F., 2011. A survey on question answering technology from an information retrieval perspective. Inform. Sci. 181 (24), 5412–5434.

Kumari, S., Pushphavati, T., 2022. Question answering and text generation using BERT and GPT-2 model. In: Computational Methods and Data Engineering: Proceedings of ICCMDE 2021. Springer, pp. 93–110.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S., 2020. A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. 30, 121–204.

Kwok, C.C., Etzioni, O., Weld, D.S., 2001. Scaling question answering to the web. In: Proceedings of the 10th International Conference on World Wide Web. pp. 150–161.

Lao, M., Guo, Y., Pu, N., Chen, W., Liu, Y., Lew, M.S., 2021. Multi-stage hybrid embedding fusion network for visual question answering. Neurocomputing 423, 541–550.

Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H., 2006. Beyond information retrieval—medical question answering. In: AMIA Annual Symposium Proceedings. Vol. 2006, American Medical Informatics Association, p. 469.

Lee, K., Salant, S., Kwiatkowski, T., Parikh, A., Das, D., Berant, J., 2016. Learning recurrent span representations for extractive question answering. arXiv preprint arXiv:1611.01436.

Lee, H., Yoon, S., Dernoncourt, F., Kim, D.S., Bui, T., Shin, J., Jung, K., 2020. KPQA: A metric for generative question answering using keyphrase weights. arXiv preprint arXiv:2005.00192.

Lee, J., Yun, S., Kim, H., Ko, M., Kang, J., 2018. Ranking paragraphs for improving answer recall in open-domain question answering. arXiv preprint arXiv:1810.00494.

Lewis, M., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Lewis, P., Denoyer, L., Riedel, S., 2019a. Unsupervised question answering by cloze translation. arXiv preprint arXiv:1906.04980.

Lewis, M., Fan, A., 2018. Generative question answering: Learning to answer the whole question. In: International Conference on Learning Representations.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., Schwenk, H., 2019b. MLQA: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474.

Li, Q., Fu, J., Yu, D., Mei, T., Luo, J., 2018. Tell-and-answer: Towards explainable visual question answering using attributes and captions. arXiv preprint arXiv:1801.09041.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. Neural Netw. Learn. Syst. 33 (12), 6999–7019.

Li, H., Wang, Y., De Melo, G., Tu, C., Chen, B., 2017. Multimodal question answering over structured data with ambiguous entities. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 79–88.

Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.-R., Hu, D., 2022. Learning to answer questions in dynamic audio-visual scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19108–19118.

Li, Y., Yang, T., 2018. Word embedding for understanding natural language: a survey. Guid. Big Data Appl. 83–104.

Li, Y., Yang, N., Wang, L., Wei, F., Li, W., 2023. Generative retrieval for conversational question answering. Inf. Process. Manage. 60 (5), 103475.

Liakata, M., Dobnik, S., Saha, S., Batchelor, C., Schuhmann, D.R., 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 747–757.

Liang, W., Jiang, Y., Liu, Z., 2021. GraghVQA: Language-guided graph neural networks for graph-based visual question answering. arXiv preprint arXiv:2104.10283.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, Y., Chowdhury, S., Zhang, C., Caragea, C., Yu, P.S., 2020a. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. arXiv preprint arXiv:2008.02434.

Liu, B., Wei, H., Niu, D., Chen, H., He, Y., 2020b. Asking questions the human way: Scalable question-answer generation from text corpus. In: Proceedings of the Web Conference 2020. pp. 2032–2043.

Liu, S., Yu, C., Meng, W., 2005. Word sense disambiguation in queries. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. pp. 525–532.

Liu, Y., Zhang, C., Yan, X., Chang, Y., Yu, P.S., 2019. Generative question refinement with deep reinforcement learning in retrieval-based QA system. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 1643–1652.

Liu, Y., Zhang, X., Zhang, Q., Li, C., Huang, F., Tang, X., Li, Z., 2021. Dual self-attention with co-attention networks for visual question answering. Pattern Recognit. 117, 107956.

Louis, A., van Dijck, G., Spanakis, G., 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38, pp. 22266–22275.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inf. Process. Syst. 32.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., Zheng, W., 2023. The multi-modal fusion in visual question answering: a review of attention mechanisms. PeerJ Comput. Sci. 9, e1400.

Lukovnikov, D., Fischer, A., Lehmann, J., Auer, S., 2017. Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1211–1220.

Ly, A., Uthayasooriyar, B., Wang, T., 2020. A survey on natural language processing (nlp) and applications in insurance. arXiv preprint arXiv:2010.00462.

Ma, Z., Sun, A., Yuan, Q., Cong, G., 2015. A tri-role topic model for domain-specific question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29.

Malinowski, M., Fritz, M., 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Adv. Neural Inf. Process. Syst. 27.

Manjunatha, V., Saini, N., Davis, L.S., 2019. Explicit bias discovery in visual question answering models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9562–9571.

Manmadhan, S., Kovoor, B.C., 2020. Visual question answering: a state-of-the-art review. Artif. Intell. Rev. 53 (8), 5705–5745.

Manna, R., Das, D., Gelbukh, A., 2020. Information retrieval-based question answering system on foods and recipes. In: Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020, Proceedings, Part II 19. Springer, pp. 260–270.

Manning, C.D., 2008. Introduction to information retrieval.

Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., Chen, W., 2020. Generation-augmented retrieval for open-domain question answering. arXiv preprint arXiv:2009.08553.

Masalkhi, M., Ong, J., Waisberg, E., Lee, A.G., 2024. Google DeepMind's gemini AI versus ChatGPT: A comparative analysis in ophthalmology. Eye 1–6.

Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E., 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244.

Maulud, D.H., Zeebaree, S.R., Jacksi, K., Sadeeq, M.A.M., Sharif, K.H., 2021. State of art for semantic analysis of natural language processing. Qubahan Acad. J. 1 (2), 21–28.

Min, S., Seo, M., Hajishirzi, H., 2017. Question answering through transfer learning from large fine-grained supervision data. arXiv preprint arXiv:1702.02171.

Mitra, M., Chaudhuri, B., 2000. Information retrieval from documents: A survey. Inf. Retr. 2, 141–163.

Mitra, B., Simon, G., Gao, J., Craswell, N., Deng, L., 2016. A proposal for evaluating answer distillation from web data. In: Proceedings of the SIGIR 2016 WebQA Workshop. Vol. 230.

Moholkar, K., Patil, S., 2021. Deep ensemble approach for question answer system. In: Computer Networks, Big Data and IoT: Proceedings of ICCBI 2020. Springer, pp. 15–24.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investig. 30 (1), 3–26.
Nararatwong, R., Kertkeidkachorn, N., Ichise, R., 2022. KIQA: Knowledge-infused question answering model for financial table-text data. In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. pp. 53–61.
Nguyen, V.-T., Le, A.-C., 2018. Deep neural network-based models for ranking question-answering pairs in community question answering systems. In: Integrated Uncertainty in Knowledge Modelling and Decision Making: 6th International Symposium, IUKM 2018, Hanoi, Vietnam, March 15-17, 2018, Proceedings 6. Springer, pp. 179–190.
Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L., 2016. Ms marco: A human-generated machine reading comprehension dataset.
Niu, Y., Huang, F., Liang, J., Chen, W., Zhu, X., Huang, M., 2021. A semantic-based method for unsupervised commonsense question answering. arXiv preprint arXiv:2105.14781.
Nordhoff, S., Lee, J.D., Calvert, S.C., Berge, S., Hagenzieker, M., Happee, R., 2023. (Mis-) use of standard autopilot and full self-driving (FSD) beta: Results from interviews with users of Tesla's FSD Beta. Front. Psychol. 14, 1101520.
Nourali, K., Dolkhani, E., 2024. Scene text visual question answering by using YOLO and STN. Int. J. Speech Technol. 27 (1), 69–76.
Olawade, D.B., Wada, O.Z., Odetayo, A., David-Olawade, A.C., Asaolu, F., Eberhardt, J., 2024. Enhancing mental health with artificial intelligence: Current trends and future prospects. J. Med. Surg. Public Heal. 100099.
Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.-J., Hiszpanski, A.M., 2020. Data-driven materials research enabled by natural language processing and information extraction. Appl. Phys. Rev. 7 (4).
Osman, A., Samek, W., 2019. DRAU: dual recurrent attention units for visual question answering. Comput. Vis. Image Underst. 185, 24–30.
Ostrovsky, R., Skeith III, W.E., 2007. A survey of single-database private information retrieval: Techniques and applications. In: International Workshop on Public Key Cryptography. Springer, pp. 393–411.
Pakray, P., Bhaskar, P., Banerjee, S., Pal, B.C., Bandyopadhyay, S., Gelbukh, A.F., 2011. A hybrid question answering system based on information retrieval and answer validation. In: CLEF (Notebook Papers/Labs/Workshop). Vol. 96.
Palangi, H., Smolensky, P., He, X., Deng, L., 2018. Question-answering with grammatically-interpretable representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32.
Pandya, H.A., Bhatt, B.S., 2021. Question answering survey: Directions, challenges, datasets, evaluation matrices. arXiv preprint arXiv:2112.03572.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.
Park, J., Cho, Y., Lee, H., Choo, J., Choi, E., 2021. Knowledge graph-based question answering with electronic health records. In: Machine Learning for Healthcare Conference. PMLR, pp. 36–53.
Pearce, K., Zhan, T., Komanduri, A., Zhan, J., 2021. A comparative study of transformer-based language models on extractive question answering. arXiv preprint arXiv:2110.03142.
Perera, N., Priyankara, C., Jayasekara, D., 2020. Identifying irrelevant answers in web based question answering systems. In: 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, pp. 11–16.
Pezeshkpour, P., 2023. Measuring and modifying factual knowledge in large language models. arXiv preprint arXiv:2306.06264.
Putra, S.J., Gusmita, R.H., Hulliyah, K., Sukmana, H.T., 2016. A semantic-based question answering system for Indonesian translation of quran. In: Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services. pp. 504–507.
Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., Wang, H., 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2010.08191.
Qu, C., Yang, L., Qiu, M., Croft, W.B., Zhang, Y., Iyyer, M., 2019. BERT with history answer embedding for conversational question answering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1133–1136.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67.
Raiman, J., Miller, J., 2017. Globally normalized reader. arXiv preprint arXiv:1709.02828.
Rajpurkar, P., Jia, R., Liang, P., 2018. Know what you don't know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822.
Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P., 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
Rakotoson, L., Letaillieur, C., Massip, S., Laleye, F.A., 2022. Extractive-boolean question answering for scientific fact checking. In: Proceedings of the 1st International Workshop on Multimedia AI Against Disinformation. pp. 27–34.
Ramakrishnan, S.K., Pal, A., Sharma, G., Mittal, A., 2017. An empirical evaluation of visual question answering for novel objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4392–4401.
Ramesh, K., Ravishankaran, S., Joshi, A., Chandrasekaran, K., 2017. A survey of design techniques for conversational agents. In: International Conference on Information, Communication and Computing Technology. Springer, pp. 336–350.
Ravi, S., Chinchure, A., Sigal, L., Liao, R., Shwartz, V., 2023. Vlc-bert: Visual question answering with contextualized commonsense knowledge. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1155–1165.
Reddy, S., Chen, D., Manning, C.D., 2019. Coqa: A conversational question answering challenge. Trans. Assoc. Comput. Linguist. 7, 249–266.
Ren, P., Chen, Z., Ren, Z., Kanoulas, E., Monz, C., De Rijke, M., 2021. Conversations with search engines: SERP-based conversational response generation. ACM Trans. Inf. Syst. (TOIS) 39 (4), 1–29.
Ren, M., Kiros, R., Zemel, R., 2015. Exploring models and data for image question answering. Adv. Neural Inf. Process. Syst. 28.
Ribeiro, M.T., Guestrin, C., Singh, S., 2019. Are red roses red? evaluating consistency of question-answering models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6174–6184.
Riloff, E., Thelen, M., 2000. A rule-based question answering system for reading comprehension tests. In: ANLP-NAACL 2000 Workshop: Reading Comprehension Tests As Evaluation for Computer-Based Language Understanding Systems.
Robertson, S., Zaragoza, H., et al., 2009. The probabilistic relevance framework: BM25 and beyond. Found. Trends® Inf. Retr. 3 (4), 333–389.
Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiskỳ, T., Blunsom, P., 2015. Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664.
Saggion, H., 2004. Identifying definitions in text collections for question answering. In: LREC.
Saha, A., Pahuja, V., Khapra, M., Sankaranarayanan, K., Chandar, S., 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32.
Sarrouti, M., El Alaoui, S.O., 2017. A machine learning-based method for question type classification in biomedical question answering. Methods Inf. Med. 56 (03), 209–216.
Savery, M., Abacha, A.B., Gayen, S., Demner-Fushman, D., 2020. Question-driven summarization of answers to consumer health questions. Sci. Data 7 (1), 322.
Seo, M., Kwiatkowski, T., Parikh, A.P., Farhadi, A., Hajishirzi, H., 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. arXiv preprint arXiv:1804.07726.
Seo, M., Lee, J., Kwiatkowski, T., Parikh, A.P., Farhadi, A., Hajishirzi, H., 2019. Real-time open-domain question answering with dense-sparse phrase index. arXiv preprint arXiv:1906.05807.
Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y., 2016a. Generating factoid questions with recurrent neural networks: The 30 m factoid question-answer corpus. arXiv preprint arXiv:1603.06807.
Serban, I., Sordoni, A., Bengio, Y., Courville, A., Pineau, J., 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 30.
Shah, M., Chen, X., Rohrbach, M., Parikh, D., 2019a. Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6649–6658.
Shah, A.A., Ravana, S.D., Hamid, S., Ismail, M.A., 2019b. Accuracy evaluation of methods and techniques in web-based question answering systems: a survey. Knowl. Inf. Syst. 58, 611–650.
Sharma, H., Jalal, A.S., 2021. A survey of methods, datasets and evaluation metrics for visual question answering. Image Vis. Comput. 116, 104327.
Sheker, M., Saad, S., Abood, R., Shakir, M., 2016. Domain-specific ontology-based approach for Arabic question answering. J. Theor. Appl. Inf. Technol. 83 (1), 43.
Shen, L., Meng, F., Zhang, J., Feng, Y., Zhou, J., 2021. GTM: A generative triple-wise model for conversational question generation. arXiv preprint arXiv:2106.03635.
Shen, W., Wang, J., Han, J., 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. 27 (2), 443–460.
Silva, J., Nogueira, P., Martins, P., Vaz, P., Abrantes, J., 2023. Exploring the potential of mixed reality as a support tool for industrial training and production processes: A case study utilizing Microsoft HoloLens. In: International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence. Springer, pp. 187–196.
Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H., 2024. Drivelm: Driving with graph visual question answering. In: European Conference on Computer Vision. Springer, pp. 256–274.
Singh, H., Nasery, A., Mehta, D., Agarwal, A., Lamba, J., Srinivasan, B.V., 2021a. Mimoqa: Multimodal input multimodal output question answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5317–5332.
Singh, D., Reddy, S., Hamilton, W., Dyer, C., Yogatama, D., 2021b. End-to-end training of multi-document reader and retriever for open-domain question answering. Adv. Neural Inf. Process. Syst. 34, 25968–25981.
Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023. Large language models encode clinical knowledge. Nature 620 (7972), 172–180.

So, B., Byun, K., Kang, K., Cho, S., 2022. Jaquad: Japanese question answering dataset for machine reading comprehension. arXiv preprint arXiv:2202.01764.

Soares, M.A.C., Parreiras, F.S., 2020. A literature review on question answering techniques, paradigms and systems. J. King Saud Univ.- Comput. Inf. Sci. 32 (6), 635–646.

Song, L., Wang, Z., Hamza, W., 2017. A unified query-based generative model for question generation and question answering. arXiv preprint arXiv:1709.01058.

Soricut, R., Brill, E., 2006. Automatic question answering using the web: Beyond the factoid. Inf. Retr. 9, 191–206.

Spennemann, D.H., 2023. ChatGPT and the generation of digitally born "knowledge": How does a generative AI language model interpret cultural heritage values? Knowledge 3 (3), 480–512.

Srihari, R., Li, W., Li, X., 1999. Information extraction supported question answering. In: TREC.

Sriramanan, G., Bharti, S., Sadasivan, V.S., Saha, S., Kattakinda, P., Feizi, S., 2024. Llm-check: Investigating detection of hallucinations in large language models. Adv. Neural Inf. Process. Syst. 37, 34188–34216.

Sufi, F., 2024. Addressing data scarcity in the medical domain: A GPT-based approach for synthetic data generation and feature extraction. Information 15 (5), 264.

Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., Ren, Z., 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542.

Tahsin Mayeesha, T., Md Sarwar, A., Rahman, R.M., 2021. Deep learning based question answering system in Bengali. J. Inf. Telecommun. 5 (2), 145–178.

Talpes, E., Sarma, D.D., Venkataramanan, G., Bannon, P., McGee, B., Floering, B., Jalote, A., Hsiong, C., Arora, S., Gorti, A., et al., 2020. Compute solution for tesla's full self-driving computer. IEEE Micro 40 (2), 25–35.

Tang, D., Duan, N., Qin, T., Yan, Z., Zhou, M., 2017. Question answering and question generation as dual tasks. arXiv preprint arXiv:1706.02027.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G., 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. pp. 41–47.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., Manning, C.D., 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint arXiv:2305.14975.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Towell, G., Voorhees, E.M., 1998. Disambiguating highly ambiguous words. Comput. Linguist. 24 (1), 125–145.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al., 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16, 1–28.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., Krahmer, E., 2019. Best practices for the human evaluation of automatically generated text. In: Proceedings of the 12th International Conference on Natural Language Generation. pp. 355–368.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575.

Voorhees, E.M., Harman, D.K., et al., 2005. TREC: Experiment and Evaluation in Information Retrieval, vol. 63, Citeseer.

Voorhees, E.M., Tice, D.M., 2000. Building a question answering test collection. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 200–207.

Wang, A., Cho, K., Lewis, M., 2020a. Asking and answering questions to evaluate the factual consistency of summaries. arXiv preprint arXiv:2004.04228.

Wang, W., Feng, S., Wang, D., Zhang, Y., 2019. Answer-guided and semantic coherent question generation in open-domain conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5066–5076.

Wang, T., Huang, J., Zhang, H., Sun, Q., 2020b. Visual commonsense r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10760–10770.

Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G., 2018. QG-net: a data-driven question generation model for educational content. In: Proceedings of the Fifth Annual ACM Conference on Learning At Scale. pp. 1–10.

Wang, Y., Yasunaga, M., Ren, H., Wada, S., Leskovec, J., 2023. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21582–21592.

Wang, L., Zheng, K., Qian, L., Li, S., 2022. A survey of extractive question answering. In: 2022 International Conference on High Performance Big Data and Intelligent Systems. HDIS, IEEE, pp. 147–153.

Weld, H., Huang, X., Long, S., Poon, J., Han, S.C., 2022. A survey of joint intent detection and slot filling models in natural language understanding. ACM Comput. Surv. 55 (8), 1–38.

Weston, J., Bordes, A., Chopra, S., Rush, A.M., Van Merriënboer, B., Joulin, A., Mikolov, T., 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. ArXiv Preprint Cmp-Lg/9406033.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., Van Den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. Comput. Vis. Image Underst. 163, 21–40.

Xiong, K., Cui, A., Zhang, Z., Li, M., 2016. Neural contextual conversation learning with labeled question-answering pairs. arXiv preprint arXiv:1607.05809.

Xu, G., Rong, W., Wang, Y., Ouyang, Y., Xiong, Z., 2021. External features enriched model for biomedical question answering. BMC Bioinformatics 22 (1), 272.

Yadav, A., Singh, P.P., 2024. Closed-domain question-answering system for Indian legal acts. Int. J. Adv. Res. Comput. Sci. 15 (2).

Yang, H., De Roeck, A., Gervasi, V., Willis, A., Nuseibeh, B., 2011. Analysing anaphoric ambiguity in natural language requirements. Requir. Eng. 16, 163–189.

Yao, Y., Yang, P., Zhao, G., Yin, G., 2024. KGAgent: Learning a deep reinforced agent for keyphrase generation. IEEE/ ACM Trans. Audio Speech Lang. Process..

Yen, S.-J., Wu, Y.-C., Yang, J.-C., Lee, Y.-S., Lee, C.-J., Liu, J.-J., 2013. A support vector machine-based context-ranking model for question answering. Inform. Sci. 224, 77–87.

Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., Li, X., 2015. Neural generative question answering. arXiv preprint arXiv:1512.01337.

Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. 29 (12), 5947–5959.

Yu, J., Zhang, W., Lu, Y., Qin, Z., Hu, Y., Tan, J., Wu, Q., 2020. Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. IEEE Trans. Multimed. 22 (12), 3196–3209.

Zellers, R., Bisk, Y., Schwartz, R., Choi, Y., 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326.

Zhang, R., Guo, J., Chen, L., Fan, Y., Cheng, X., 2021. A review on question generation from natural language text. ACM Trans. Inf. Syst. (TOIS) 40 (1), 1–43.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al., 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Zhao, T., Yan, Z., Cao, Y., Li, Z., 2021. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 3948–3954.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., Han, J., 2022. Towards a unified multi-dimensional evaluator for text generation. arXiv preprint arXiv:2210.07197.

Zhu, B., Jordan, M., Jiao, J., 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In: International Conference on Machine Learning. PMLR, pp. 43037–43067.

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.-S., 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774.