Journal of Computing and Security

Research Article, 2025, Volume 12, Number 2 (pp. 49–76)

http://www.jcomsec.org

A Comprehensive Survey on Multi-hop Machine Reading Comprehension Datasets and Metrics

Azade Mohammadi [a] Reza Ramezani [a,*]

[a] *Computer Engineering Faculty,University of Isfahan, Isfahan, Iran.*

ARTICLE INFO.

Article history:
Received: 12 November 2024
Revised: 03 June 2025
Accepted: 30 June 2025
Published Online: 20 December 2025

Keywords:
Multi-hop Machine Reading Comprehension, Multi-hop Machine Reading Comprehension Dataset, Natural Language Processing.

ABSTRACT

Multi-hop Machine reading comprehension is a challenge field in natural language processing and its goal is to find the answer of questions based on disjoint pieces of information in a natural language context. Due to the complexity and importance of this field, a large number of studies have been focused on multi-hop MRC recently. Then it is necessary and worth reviewing them in detail. Since datasets and evaluation metrics play a crucial role in the progress of research in this field, this study aims to present a comprehensive survey on recent advances in multi-hop MRC evaluation metrics and datasets. In this regard, the existing multi-hop MRC evaluation metrics and datasets are reviewed from 2017 to 2025, and a comprehensive analysis has been prepared at the end and. Also, open issues in this field have been discussed.

1 Introduction

Machine reading comprehension (MRC) is one of the most important and long-standing topics in Natural Language Processing (NLP). MRC provides a way to evaluate an NLP system's capability for natural language understanding. An MRC task, in brief, refers to the ability of a computer to read and understand natural language context and then find the answer to questions about that context. The emergence of large-scale single-document MRC datasets, such as SQuAD[1] , CNN/Daily mail[2], has led to increased attention to this topic and different models have been proposed to address the MRC problem, such as [3],[4], [5] and [6]. However, for many of these datasets, it has been found that models don't need to comprehend and reason to answer a question. For example, Khashabi et al [7] proved that adversarial perturbation in candidate answers has a negative effect on the performance of the QA systems. Similarly, [8] showed that adding an adversarial sentence to the SQuAD [1] context will drop the result of many existing models. Also [3] pointed out that the required reasoning in the CNN/Daily Mail [2] dataset is so simple that even a relatively simple algorithm can perform well on this dataset . Min et al. [9] have shown that 90% of the questions in SQuAD [1], are answerable given only one sentence in a document, which does not involve complex reasoning. The above problems seem to be due to the fact that answering the questions of these datasets doesn't require a deep understanding and reasoning. [7]. In other words, these datasets have focused only on answering questions based on a single or few nearby sentences of the context, mostly by matching information in the question and the context (known as single-hop MRC). However, in real-world cases, to answer a question it is required to read and comprehend multiple parts of disjoint evidence to find the valid information. Therefore, there are gaps between single-hop MRC, and real-world cases. Multi-Hop Reading Comprehension (MHRC)

* Corresponding author.
Email addresses: azade.mohammadi@eng.ui.ac.ir (A. Mohammadi), r.ramezani@eng.ui.ac.ir (R. Ramezani)
https://dx.doi.org/10.22108/jcs.2025.143323.1153
ISSN: 2322-4460

is a more challenging extension of MRC in which it is needed to properly integrate multiple pieces of evidence and reason over them to correctly answer a question. In contrast to the question in single-hop MRC that can be answered by matching information, the multi-hop MRC task requires answering more complex questions based on a deep understanding of the full information. The reasoning ability is considered the first key in multi-hop MRC [10] . Figure 1 shows two MRC triples. Figure 1(a) shows a single-hop question from [1] that as you can see this question can be answered with one sentence and mostly used matching information in the question and the context (precipitation, fall). Figure 1(b) shows an example from the WikiHop dataset [11], where the machine has to gather information from three passages and reason over them to choose the correct answer among candidate answers. Passages contain two relevant facts: "The Hanging Gardens are in Mumbai" and "Mumbai is a city in India" and also some irrelevant facts, like: "The Hanging Gardens provide sunset views over the Arabian Sea" and "The Arabian Sea is bounded by Pakistan and Iran". As it is clear, finding the final answer in this kind of scenario is more challenging.

The first attempt to improve the simple single-hop MRC dataset happened with emerging of some datasets like TriviaQA [12] and NarrativeQA [13]. These datasets impose more challenges by introducing multiple passages per each question, and also presenting the questions that couldn't be answered with one single sentence. Although they present multiple passages per each question, still the most portion of the question in these datasets could be answered by exactly one sentence or a few nearby sentences within one passage, which means they do not need multi-hop reasoning for most percent of the questions. They are generally known as multi-passage or multi-document dataset that is closer to open-domain Question Answering or retrieving-reading problem, which means models have to focus on retrieving the most related passage and then answer the question based on that instead of reasoning over disjoint information. HotpotQA [14] and WikiHop [11] can be mentioned as the first and most popular multi-hop datasets which in addition to providing multiple passages per each question, ensure that the question can only be answered by reasoning over disjoint pieces of information. It has been shown that the models with successful results in single-hop MRC datasets have limited success on these datasets [15]. The advancement of MRC metrics from single-hop to multi-hop tasks has involved a shift from primarily focusing on answer span overlap (like Exact Match and F1 score) to incorporating measures that can assess the model's ability to identify and utilize the multiple pieces of evidence required for reasoning. While standard metrics are still used, multi-hop MRC evaluation increasingly emphasizes the need for models to not only provide the correct answer but also to identify the supporting facts or reasoning paths. This has led to the development of metrics that evaluate the selection of these supporting sentences, either as a separate task or integrated into the overall evaluation, aiming to better reflect the true multi-hop reasoning capability of the models. However, developing comprehensive metrics that fully capture the complexity of multi-hop inference and the plausibility of the reasoning remains an ongoing challenge in the field. Recently, a lot of studies have been focused on multi-hop MRC, and each of them focuses on one different aspect of this task. Datasets are a vital part of the MRC task because, without a proper dataset, it is not possible to train or evaluate the models. It can be claimed that the emergence of datasets is the main reason for much attention and progress in the MRC field after 2016. (Chen, 2018). Usually, the aim of each new dataset is to propose some new challenges that have been neglected by previous datasets. Since new challenges have not been addressed by the existing models, they usually can't achieve the acceptable result in new datasets, and then the models needed to be improved. Therefore, it can be said that the high-quality multi-hop datasets promote the multi-hop models as well. To have an accurate investigation of the growing trend of creating the multi-hop dataset, we need to consider the dataset alongside proposing multi-hop models, which the number of them has been shown in Figure 2. In 2017, several datasets were introduced (like TriviaQA [12]) that are not considered multi-hop datasets, but they attracted attention to this task by proposing more complicated questions than single-hop datasets. Proposing multi-hop models has been beginning from 2018, but there was a shortage of real multi-hop datasets, so in 2018 some MRC datasets were been proposed with a strong focus on the multi-hop challenges. These datasets made a proper situation to present the multi-hop MRC models, as you can see a significant number of models were being proposed in 2019. The trend of introducing new multi-hop datasets continued from 2020 to 2025. It can be inferred from Figure 2 that whenever a new dataset has been created, many new models will be proposed after that to address new presented challenges, which shows the importance of the datasets.

There are some review papers on the MRC task that we will mention here to explain the difference and the innovation aspect of our paper. Liu et al. [16] reviewed 85 studies from 2015 to 2018 with a focus on neural network solutions for the MRC problem

In meteorology, **precipitation** is any product of the condensation of atmospheric water vapor that **falls** under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called \showers".	[The Hanging Hardens], in [Mumbai], also known as Pherozshah Mehta Garden, are terraced garden… [They] provide sunset views over [Arabian Sea].. [Mumbai] (also known as Bombay), the official name until 1950 is the capital city of the Indian state of Maharashtra. [It] is the most populous city in [India]… The [Arabian Sea] is a region and [Iran], on the west by northeastern [Somalia] and the Arabian Peninsula, and on the east by ….
Q: What causes precipitation to fall?	Q: (The Hanging Garden, Country, ?) Options: Iran, India, Pakistan, Somalia, …
Answer: gravity	Answer: India

Figure 1. An Example of Single-Hop and Multi-Hop Question.

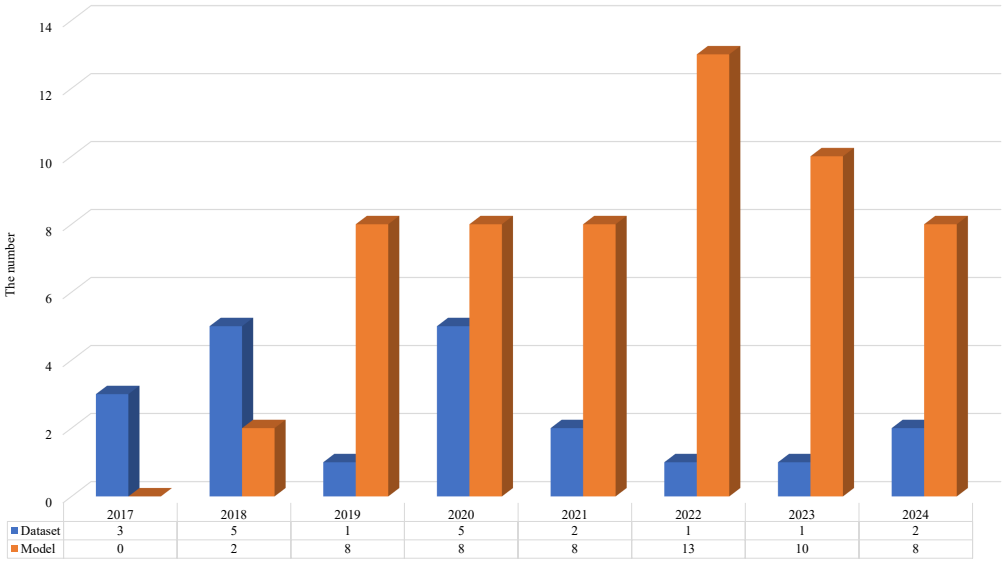


Figure 2. The Number of the Multi-Hop Models and Datasets in Recent Years.

to investigate the neural network methods in the MRC task. Baradaran, Ghiasi and Amirkhani [17] presented a survey on the MRC field based on 124 reviewed papers from 2016 to 2018 with a focus on presenting a comprehensive survey on different aspects of machine reading comprehension systems, including their approaches, structures, input/outputs, and research novelties. Thayaparan, Valentino and Freitas [18] proposed a systematic review about explainable MRC, from 2014 to 2020 with a focus on the explainable feature of the recent MRC methods. Zhang, Zhao and Wang [19] presented a survey on the role of contextualized language models (CLMs) on MRC from 2015 to 2019. Bai and Wang [20] presented a survey on textual question answering with a focus on datasets and metrics, they investigate 47 datasets and 8 metrics.

Although these surveys investigate a different aspect of MRC/QA, none of them have focused on the multi-hop challenges. Due to the importance and increasing attention to multi-hop MRC it is necessary to investigate the multi-hop MRC studies separately. Our contribution in this paper is to focus on multi-hop MRC evaluation metrics and datasets with a focus on their multi-hop challenges.

A survey focused on multi-hop MRC provides researchers and enthusiasts interested in the challenges of multi-step reasoning with deeper, more relevant, and more coherent information, helping them to better understand this important and growing subfield.

Then this study will:

- Provide a clearer picture of the current state of research in this subfield.
- Identify open problems and future research directions more precisely.
- Offer a focused collection of key resources and datasets relevant to their specific area of interest.

In this regard, we first introduce the problem definition of the multi-hop MRC task and how it differs from general MRC. Also, investigate the evaluation metrics and the unique features of them to be appropriate for multi-hop MRC tasks, then review 23 multi-hop MRC datasets from 2017 to 2025. Also, a fine-grain analysis of the datasets will be presented with a focus on different aspects. Finally, open issues in this field have been discussed. It is important to note that since there is a close relationship between MRC and Question Answering, most of the existing machine reading comprehension tasks are in the form of textual question answering (Zeng et al., 2020), also MRC is known as a basic task of textual question answering [16]. Figure 3 shows the relationship between QA, MRC and multi-hop MRC [20].

The rest of this paper is organized as follows: In Section 2, the definition of the problem and different types of the multi-hop MRC task are explained. In Section 3, evaluation metrics for multi-hop MCR are investigated, Next, in section 4, existing multi-hop MRC datasets are reviewed in detail and with a focus on their multi-hop aspects. Also, a fine-grain comparison of reviewed datasets will be presented in section 5. Open issues are expressed in Section 6, and finally Section 7 concludes the paper.

2 Problem Definition

In general, the multi-hop MRC problem can be defined as: Given a collection of training examples (C;Q;A), the goal is to find a function F which takes a context C and a corresponding question Q as inputs, and gives answer A as output.

$$F : (C, Q) \rightarrow A \tag{1}$$

Since the problem is multi-hop, $C = (P_1, P_2, , P_lp)$ can be a set of paragraphs (or documents) where lp denotes the number of paragraphs (or documents) and also question Q is such a way that needs the multiple disjoint pieces of information from C to be answered. Like general MRC task, Answer A in multi-hop MRC can be in different forms, where have been divided into four categories [21]:

- Span-extraction: The span extraction task needs to extract the subsequence A from Pi ($P(i) \in C$) as the correct answer of question Q by learning the function F, such that A=F(C,Q).
- Multiple-choice: Given a set of candidate answers $A = A_1, A_2, , A_n$, the multiple-choice task needs to select the correct answer Ai from n possible answer by learning the function F, such that Ai=F(C,Q).
- Free-form: The correct answer is A that $A \in C$ or $A \notin C$. In other words, the answer is not necessarily limited to be a part of the passage. The free-form task needs to predict the correct answer A by learning the function F, such that A=F(C,Q).
- Cloze-style: The correct answer A is part of the question Q (usually a word or an entity) that is removed from question. The cloze style task needs to fill in the blank with the correct word or entity A by learning the function F, such that A=F(C,Q).

All General MRC tasks have the potential to be used in form of multi-hop MRC form as well. But as Figure 4 shows, most existing multi-hop datasets focus on the first two tasks (span-extraction and multiple-choice). it should be said that paying attention to the two first tasks is not specific for multi-hop and it happens for general MRC too. The great attention to these tasks may be because they are more natural and similar to the real-world case in comparison to the cloze-style task. Although the free-from task is more similar to the real-world case among all tasks, it didn't achieve proper attention which can be because of the complexity of this task for creating datasets and proposing models.

3 Evaluation Metrics

Evaluation metrics play a crucial role in multi-hop machine reading comprehension (MRC) by providing a quantitative way to assess how well a model can reason across multiple pieces of text to answer a question. Unlike single-hop MRC where the answer is typically found within a single context passage, multi-hop MRC requires the model to integrate information from several disjointed text segments. This complexity necessitates evaluation metrics that go beyond simple answer extraction and consider the reasoning process. In this section, the evaluation metrics used in the multi-hop MRC task will be reviewed and investigated.

Exact Match: This metric is used to show whether two texts (predicated answer and the ground truth) are exactly the same. If they are exactly the same, EM will be 1, otherwise, it will be 0. This metric is a pop-

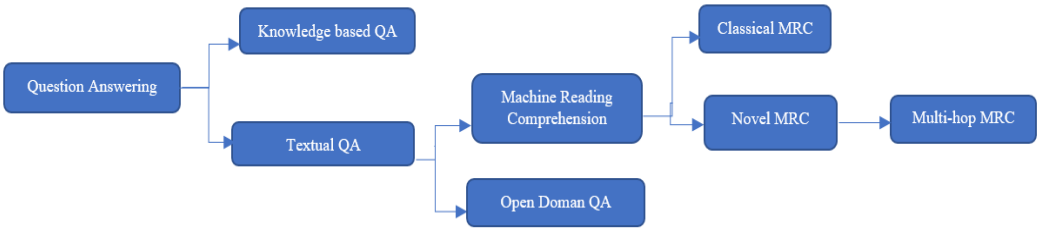


Figure 3. The Relationship Between QA and MRC.

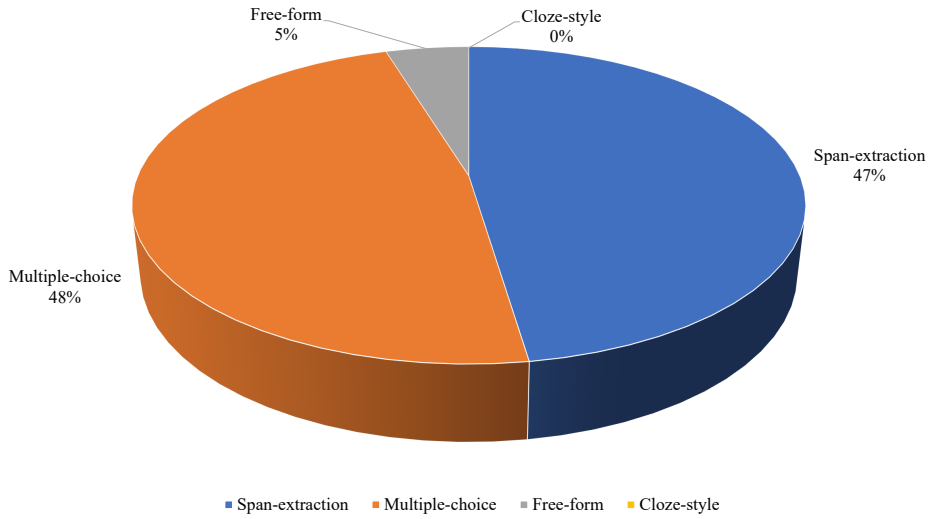


Figure 4. The Frequency of Multi-Hop MRC Tasks in Reviewed Datasets.

ular and useful metric in the span-extraction task, as in this task the predicated answer should be exactly matched with the ground-truth answer. Even minor discrepancies like an extra word, a missing punctuation mark, or a synonym result in a score of 0. This can underestimate the performance of models that provide answers that are semantically correct but not lexically identical to the reference. EM does not account for semantic similarity between the predicted and ground truth answers. Two completely different phrasings that mean the same thing will both be penalized if they don't exactly match the reference. EM becomes increasingly difficult to achieve for longer, more descriptive answers where slight variations in wording are common and acceptable. F1 Score: The F1 score metric measures the average overlap between the predicted answer and the ground-truth one which is widely used in the span-extraction task. It is calculated with precision and recall measurements, that we briefly show as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1 = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{4}$$

where TP is True Positive, FP is False Positive, TN is True Negative and FN is False Negative. Consider the predicated and ground-truth answers as a bag of tokens. Accordingly, the definition of these parameter is shown in Table 1.

F1 score, which calculates the harmonic mean of precision and recall at the token level, is more lenient than EM. However, it still relies on lexical overlap. While it can capture partial correctness (e.g., if a predicted answer shares some words with the ground truth), it doesn't fully address semantic understanding. Two

Table 1. TP, FP, TN and FN [16].

	Tokens in ground-truth	**Tokens not in ground-truth**
Tokens in predicated answer	TP	FP
Tokens not in the predicated answer	FN	TN

previous metrics are common and popular in many NLP tasks as well as general MRC. There are three sets of metrics based on them that are more common and important in multi-hop MRC/QA models that have been introduced by Yang et al. [14]. The main goal of multi-hop MRC models is to find the answer then EM and F1 are used to evaluate this task known as the Answer task. Also, multi-hop MRC should present a series of evidence sentences that have led to the final answer known as the Support task as proof of multi-hop reasoning, this task is for evidence sentence extraction and could be seen as a kind of reasoning interpretability (Wu et al., 2021). Also, there is a task to evaluate both above tasks known as the Joint task. Each metric is explained in detail as follows:

Answer F1: It focusses on the answer which is a span of the context. It measures the average overlap between the predicted answer span and the ground-truth one.

Answer EM: It focuses on the answer which is a span of the context. It is used to show whether the predicated answer span and the ground-truth are exactly the same.

Support F1: it focuses on the supporting fact and to show the expansibility of the Multi-hop MRC models. It measures the average overlap between of the supporting facts compared to the gold set.

Support EM: It focuses on the supporting facts and to show the expansibility of the Multi-hop MRC models. It is used to show whether the supporting fact set and the gold set are exactly the same.

Joint F1: it is to combine the evaluation of answer span and supporting facts: given precision and recall on the answer span (P_{ans}, R_{ans}) and the supporting facts P_{sup},R_{sup}, respectively, joint F1 is calculated as [14]:

$$P_{joint} = P_{ans}P_{sup}, R_{joint} = R_{ans}R_{sup} \tag{5}$$

$$jointF1 = \frac{2P_joint R_joint}{P_joint + R_joint} \tag{6}$$

Joint EM: is to combine the evaluation of answer span and supporting facts. It is 1 only if both tasks achieve an exact match and otherwise 0.

Accuracy: Accuracy is a popular and fairly common metric to evaluate the performance of the multiple-choice and cloze-style MRC tasks. In the multiple-choice task, it is required to check whether the correct answer has been selected from the candidate answers, and in the Cloze-style task, it is required to check whether the correct words have been selected for the missing words. The accuracy metric is calculated as follows:

$$Accuracy = \frac{N}{M} \tag{7}$$

where N is the number of questions answered correctly, and M is the number of all queried questions. Unlike EM and F1 score, there is no any supporting fact version of accuracy and it is only used for the answer prediction task.

BLEU: The BLEU (BiLingual Evaluation Understudy) metric was firstly proposed by Papineni et al. [22] for machine translation, but it can be easily used in the free-form MRC task too. BLEU is used to calculate the similarity of two sentences. In the free-form MRC task, the predicted answer and the ground truth one is those two sentences whose similarity should be calculated. In this case, the probability of n-grams in the candidate sentence that appear in the ground truth will be calculated [16]:

$$p_n = \frac{\sum_i \sum_k min(h_k(c_i), max(h_k(r_k)}{\sum_i \sum_k (h_k)(c_i)} \tag{8}$$

Where $h_k(c_i)$ counts the number of k^{th}n-gram appearing in the candidate answer c_i. In a similar way, $h_k(r_i)$ denotes the occurrence count of that n-gram in the ground-truth answer r_i. If the length of the candidate sentence is very short, the accuracy of the BLEU score will decrease. To alleviate this problem, the penalty factor BP is introduced, which can be calculated as follow [16]:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c < r \end{cases} \tag{9}$$

Finally, the BLEU is computed as:

$$BLEU = BP.exp(\sum_{(n=1)}^{N} w_n log p_n) \tag{10}$$

where N means n-grams up to length N and w_n

equals 1N. The BLEU score is the weighted average of each n-gram, and in most cases the maximum of N is 4, namely BLEU-4. **R**OUGE-L: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) has been proposed by Lin [23] and is commonly used for the machine translation and summarization tasks, but it can be used for free-form MRC as well. L in Rouge-L is longest Common Subsequence (LCS). Rouge-L means applying the longest common subsequence to measure the similarity between text and can be calculated as follows [16]:

$$R_{LCS} = \frac{LCS(X,Y)}{m} \tag{11}$$

$$P_{LCS} = \frac{LCS(X,Y)}{n} \tag{12}$$

$$F_{LCS} = \frac{(1+\beta)^2 R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{13}$$

where X is the ground-truth answer, m is the number of tokens of X,Y is the predicated answer, n is the number of tokens of Y,LCS(X,Y) is the length of the longest common subsequence of X and Y, and is a parameter to control the importance of precision P_LCS and recall R_LCS. **M**eteor: Meteor (Metric for Evaluation of Translation with Explicit ORdering) has been proposed by Banerjeeand and Lavie [24]. Like two previous metrics, it was introduced for machine translation and can be used for free-form task as well. It is claimed that it has a high level of correlation with human judgments of translation quality in comparison to BLEU. It is calculated as follows:

$$Meteor = F * (1 - Penalty) \tag{14}$$

Where F is a weighted F-score and is calculated as follows:

$$F = \frac{(Precision * Recall)}{(\alpha * Precision + (1-\alpha) * Recall)} \tag{15}$$

Also, Penalty is calculated as follows:

$$Penalty = \gamma \frac{ch}{m}^{\beta} \tag{16}$$

where ch is the total number of matching chunks and m is the total number of matched uniforms between the prediction and the ground truth. The parameters , , and are tuned to maximize the correlation with human judgments.

All three metrics correspond with the way humans would evaluate the same text and also, and they are language-independent. They have a great potential to use in multi-hop MRC, but there are some weaknesses for them. The most important one is they do not consider the meaning of the words. For example, ”watchman” and ”guard” would consider two different words. Besides they do not consider the importance of the words in the corpus for scoring. To avoid these shortcomings, following evaluation metrics have been proposed.

SMS: sentence mover's similarity used word mover's distance [25] for evaluating multi-sentence texts. It can be used to evaluate text in a continuous space using pre-trained word embedding instead of relying on exact word matching. To this end, for any two documents A and B, WMD is defined as the minimum cost of transforming one document into the other. Each document is represented by the relative frequencies of words it contains, i.e., for the ith word type:

$$d_{A,i} = \frac{count(i)}{|A|} \tag{17}$$

where $|A|$ is the total word count of document A. Also supposed that $V_i \in R^m$ is m-length embedding representation of ith words. The WMD(A,B) is calculate is follows:

$$d_{A,i} = \frac{count(i)}{|A|} \tag{18}$$

where $|A|$ is the total word count of document A. Also supposed that $V_i \in R^m$ is m-length embedding representation of ith words. The WMD(A,B) is calculate is follows:

$$WMD(A,B) = min \sum_{i=1}^{V} \sum_{j=1}^{V} T_{i,j}(i,j) \tag{19}$$

Where $\sum_{j=1}^{V} T_{i,j} = d_{A,i}$ denotes how much of word i (across all its tokens) in A is assigned to tokens of word j in B. Then WMD is transformed into a similarity (WMS) to measure two documents' similarity by minimizing the total distance to move words between two documents:

$$WMS(A,B) = exp(-WMD(A,B)) \tag{20}$$

SMS uses the WMS, except that each document is represented as a bag of sentence embedding rather than a bag of word embedding.

BERTScore: is another metric that instead of the exact words uses leverages the pre-trained contextual embedding from BERT models and matches words in candidate and reference sentences by cosine similarity. Given a reference sentence $x = x_1, , x_k$ and a candidate sentence $x = x_1, , x_l$, contextual embedding is used to represent the tokens. Then, an alignment is computed between candidate and reference words by computing pairwise cosine similarity. This alignment is then aggregated in to precision and recall scores before being aggregated into a (modified) F1 score that is weighted using inverse-document-frequency values. The cosine similarity of a reference token x_i and a candidate token x_i is calculated as:

$$\frac{x_i x_j}{||x_i||x_j||} \tag{21}$$

Note that pre-normalized vectors are used, which reduces this calculation to the inner product, while this measure considers tokens in isolation, the contextual embedding contain information from the rest of the sentence.

Figure **??** has summarized metrics and multi-hop MRC. Each metric has some features that make them suitable for one of the MRC tasks described in Section 2, then it cannot be said that one metric is better than the others as we cannot decide which one of the MRC tasks is better than the others. The frequency of usage of the evaluation metrics depends on the datasets. For example, Figure 4 showed that the number of the Span-extraction dataset is the most, then it can be concluded that EM and F1 scores are also the most used evaluation metrics which doesn't show their superiority, but it shows that there are more Span-prediction models and datasets.

4 Datasets

In this section, 23 datasets have been reviewed in detail. We choose the datasets that focus on multi-hop challenges, which means it is required to gather information and reason on multiple disjoint pieces of information to answer the question.

Datasets are sorted mainly by release date, and we also start with the simplest datasets. These datasets introduced the question which cannot be answered with a single sentence and require gathering information from more than one sentence. Although they are not considered a multi-hop dataset today, due to the simplicity of the questions in comparison to recent multi-hop questions, they can be considered the first attempt to move from simple single-hop datasets (such as SQUAD [1]) to existing multi-hop datasets, then investigating them in this section will be useful.

After a brief introduction of these simple datasets, we have investigated more complex multi-hop datasets that will be reviewed.

In the following, at first, each dataset will be introduced separately, in addition to general information, we focus on the points that make them suitable for multi-hop MRC. The end of section, contains a fine-grain comparison of reviewed datasets using several figures and tables.

NewsQA [26] contains 119,633 plain-text questions on 12,744 news articles from CNN. The answers are spans that have to be extracted from context, but the length of spans is not fixed. This dataset also contains some questions that have no answer to examine the ability of the model to recognize inadequate information. The main goal of NewsQA is to emphasize reasoning behaviors. Table 2 shows the reasoning mechanism needed to answer questions in this dataset in comparison with the SQUAD dataset [1] which is a single-hop dataset. As you can see 20.7% of questions needs synthesizing information distributed across multiple sentences, and 13.2 % of questions need to be inferred from incomplete information or by recognizing conceptual overlap. However, in 27% of questions, a single sentence in the article entails the question, and 32.7% of questions can be answered with word matching. Although the number of complex questions (3 and 4) is more than SQUAD, in comparison to the number of the other type of questions (type 1 and 2) doesn't seem enough. besides, most of the questions in type 3 can be answered with a few sentences in one single paragraph and don't need complicated reasoning.

Construction process: first the articles is retrieved from CNN/Daily Mail with the script created by Hermann et al. (2015), then a district set of crowd workers are used as follows: The first set of crowd workers (Questioners) only see a news article's headline and its summary points (also available from CNN); they do not see the full article itself. They are asked to formulate a question from this incomplete information. A second set of crowd workers provide answers. A third set of crowd workers sees the full article, a question, and the set of unique answers to that question. These workers choose the best answer from the candidate set or rejecting all answers. Each article-question pair is validated by an average of 2.48 crowd workers. Error analysis in NewsQA often reveals issues related to struggle with anaphoric resolution within the news articles, misinterpreting pronouns or other referring expressions and thus failing to identify the correct answer span.

RACE [27] has been collected from the English exams for middle and high school Chinese students and

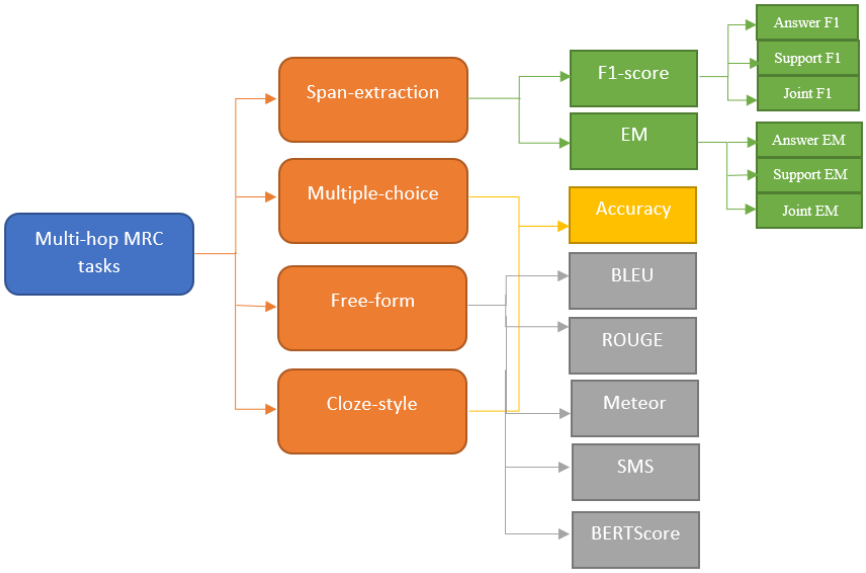


Figure 5. The Multi-hop MRC Tasks and Metrics.

Table 2. The Reasoning Mechanism [26].

	Reasoning mechanism	NewsQA	SQUAD
1	**Word matching**	32.7	39.8
2	**Paraphrasing**	27	34.3
3	**Inference**	13.2	8.6
4	**Synthesis**	20.72	11.9
5	**Ambiguous**	6.4	5.4

consists of 28,000 passages and 100,000 questions. Passages have different domains such as news, stories, ads, biography, philosophy, etc. The answers should be chosen from four candidate answers where only one of them is correct. One important point about this dataset is that the questions and answers are not limited to being from the original passage, they can be described in any words. RACE-M and RACE-H are two subgroups of RACE, where RACE-M denotes the Middle-school examinations and RACE-H denotes the High-school examinations. Some useful statistics of the RACE dataset are shown in Table 3. Also, Table 4 shows the type of reasoning mechanisms of this dataset in comparing two single-hop MRC datasets and also the NewsQA dataset (the mechanism the same as NewsQA). The number of questions from types 3 and 4 is remarkably more than CNN and SQUAD (two single-hop MRC datasets), and also is more than the NewsQA dataset, but still not enough portions of questions in RACE are in type 1(which means don't need reasoning over multiple disjoint information).

Construction process: The data collection process involved gathering a large number of passages and their associated multiple-choice questions directly from high-school exams, since real-world exam questions, providing a challenging benchmark for evaluating reading comprehension abilities on diverse textual content. Error analysis in RACE often reveals difficulties in misinterpreting nuanced language, idioms, or figurative expressions common in the exam texts.

Table 3. Statistics of the RACE Dataset [27].

Featuers	RACE-M	RACE-H	RACE
Passage len.	231.1	353.1	321.9
Question len.	9.0	10.4	10.0
Option len.	3.9	5.8	5.3
Vocab size	32811	125120	132629

TriviaQA [12] contains over 650K question-answer-evidence tuples, the context consists of the web documents and wiki documents while the older dataset like SQUAD[1] is limited to wiki documents and the News is limited to the news article. Some general statistics of TriviaQA are shown in Table 5. One of the advantages of this dataset is the syntactic and lexical variability between questions, answers, and context sentences which makes the reasoning difficult. Besides,

Table 4. The Reasoning Mechanism [27].

	Reasoning mechanism	RACE	NewsQa	CNN/Daily mail	SQUAD
1	Word matching	15.8	32.7	13	39.8
2	Paraphrasing	19.2	27	41	34.3
3	Inference	33.4	13.2	19	8.6
4	Synthesis	25.8	20.7	2	11.9
5	Ambiguous	5.8	25	6.4	5.4

there are six documents per question and also 40% of the questions need reasoning over multiple sentences, which is about three times more than SQUAD, and also is more than NewsQA [26] and RACE [27]. Figure ?? shows the percentage of the multi-sentences question in 5 datasets. However, most of these questions can be answered by a few nearby sentences in one single paragraph, and does not require complex reasoning [14]. But this dataset is still (2021) used by some multi-hop MRC models that shows the importance of this dataset.
Construction process: Trivia and quiz-league websites is used to collect question answer pairs, then questions with less than four tokens are removed to avoid vague and simple question. Also documents from the top 10 search result Web pages from Wikipedia and Bing articles is used as the textual evidence to answer questions.

Table 5. TriviaQA Statistics [12].

Total number of QA pairs	95,956
Number of unique answers	40,478
Number of evidence documents	662,659
Avg. question length (word)	14
Avg. document length (word)	2,895

COMPLEXWEBQUESTIONS [28] [28] has been built from the WEBQUESTIONS [29] and includes 34,689 question-answer pairs. This dataset is important because generating complex questions that are needed for reasoning over multiple pieces of information is the main aim challenge of this dataset, and it is claimed that all of the questions are in this form. As it has been shown in Figure ??, to generate complex questions, they used the WEBQUESTIONSSP [30], which contains 4,737 questions paired with SPARQL queries for Freebase [31], but since the questions are simple, they first generated automatically questions that are understandable to AMT workers, and then have them paraphrase those into natural language. They make sure that questions include phenomena such as composition questions (45%), conjunctions (45%), superlatives (5%), and comparatives (5%). A drawback of this method for question-generating is that because queries are generated automatically the question distribution is artificial from a semantic perspective.
Construction process: The data collection process involved formulating complex question from the dataset WEBQUESTIONSSP which contains 4,737 questions paired with SPARQL queries for Freebase and use AMT workers to paraphrase the question into the natural language form. Answering these questions necessitates interacting with a search engine by executing complex question queries against Freebase. Error analysis in ComplexWebQuestions often reveals challenges in handling the dynamic and unstructured nature of the web, including issues with information reliability and redundancy.

OpenBookQA [32] consists of 6000 plain-text questions based on 1326 elementary-level science facts. The answers are plain text which should be chosen from multiple candidate answers. Unlike other datasets that are self-contained answers, an open book fact and broad common knowledge are used together to answer the questions they require multi-hop reasoning over core facts and common knowledge to answer the question. However, it should be said that reasoning over a whole document is rarely needed actually, and also it is unclear how many additional facts are needed [33]. Some usefull statistics of OpenBookQA are shown in Table 6.
Construction process: A random science fact is shown to crowd-worker (workers from North America) to think of a second common fact, that may be combined with the first fact to derive a new, valid assertion. then the assertion is converts into a question-answer pair

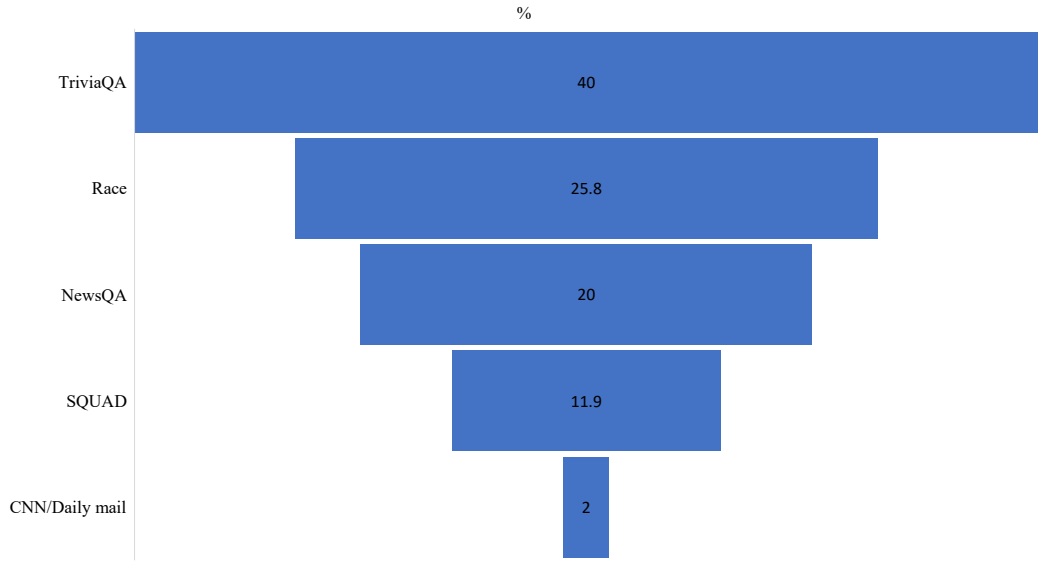


Figure 6. The Percentage of the Multi-sentences Questions.

1. Seed Question	*What movies have robert pattinson starred in?*
2. SPARQL	ns:rebert_pattinson ns:film.actor.film ?c . ?c ns:film.performance.film ?x . ?x ns:film.film.produced_by ns:erwin_stoff
3. Machine-generated	*What movies have robert pattinson starred in and that was produced by Erwin Stoff?*
4. Natural language	*Which Robert Pattinson film was produced by Erwin Stoff?*

Figure 7. Data Collection Process [28].

and extends into a 4-way multiple choice question by adding 3 incorrect answer choices. The question is also shown to 5 new crowd-workers, who are asked to answer it. If at least 4 out of 5 workers answer the question correctly, it is deemed answerable. With this construction mechanism, each question is linked to the specific core fact that support explainibility analysis, practically in the training data. Error analysis on OpenBookQA can arise from subtle linguistic variations in the questions or answer choices that mislead models relying on surface-level matching.

MultiRC ([7] et al., 2018) consists of 9k high-quality multiple-choice RC questions. It has been ensured that answering each of the questions requires reasoning over multiple sentences. This dataset is constructed using 7 different domains (news, Wikipedia articles, Articles on society, law and justice, Articles on history and anthropology, Elementary school science textbooks, 9/11 reports, and Fiction). You can see the statistics of this dataset in Table 7. Answering 60% of the questions of this dataset requires reasoning over multiple sentences. also, the required information to answer the questions is not explicitly stated in the context but is only inferable from it

Table 6. Statistics of the OpenBookQA Dataset [32].

Statistic	Value
# of questions	5,957
# of choices per question	4
Avg. question sentences	1.08 (6)
Avg. question tokens	11.46 (76)
Avg. choice tokens	2.89 (23)
Avg. science fact tokens	9.38 (28)
Vocabulary size (q+c)	11,855
Vocabulary size (q+c+f)	12,839
Answer is the longest choice	1,108 (18.6%)
Answer is the shortest choice	216 (3.6%)

(e.g., implied counts, sentiments, and relationships). The number of multi-sentence questions is 5825 (from 9,872) and the number of sentences that are used to answer a question is 2.58 on average, these sentences are not continuous and the average distance between them is 2.4. However, this dataset has focused on passage discourse and entity tracking, rather than relation composition [33].

Construction process: The data collection process involved selecting passages from various sources, including news articles, Wikipedia, and other publicly available texts. For each passage, multiple questions were created, each with several answer options. Human annotators were used to write both plausible and implausible answer choices, ensuring that the task

required genuine comprehension rather than simple keyword matching. Then, while the MultiRC dataset does not contain explanation, its nature of reasoning over multiple sentences can be used to evaluate the explainability of the models. Error analysis in MultiRC would likely focus on lacking and identifying common pitfalls like overlooking crucial information in one sentence or misinterpreting relationships between different parts of the text would be key.

Table 7. Statistical Information of MultiRC ([7] et al., 2018).

Parameter	Values
Number of Paragraphs	871
Number of Questions	9,872
Number of multi-sentence questions	5,825
Average number of candidates (per question)	5.44
Average number of correct answers (per question)	2.58
Average number of sentences used for questions	2.37
Average distance between the sentences used for each question	2.4

QAngaroo [34] is composed of two multi-hop reading comprehension datasets: MedHop which is a cloze-style dataset, and WikiHop which is a multiple-choice dataset. The MedHop dataset is about molecular biology and consists of 1.6K instances for training, 342 instances for development, and 546 instances for testing. WikiHOP dataset consists of 51k questions, answers, and context where each context consists of several documents from Wikipedia. The size of WikiHop and MedHop datasets for different learning phases are shown in Table 8. In these datasets, models have to combine evidence across multiple documents and perform multi-hop inference. WikiHop is more popular Which is probably due to being open-domain and also the number of its questions. Each question in WikiHop is a tuple, which denotes two entities, and their relationship, then the answers in the WikiHop dataset are a single entity. As Table 9 shows, for 45% of cases, the answer can be found from multiple contexts, and for 9% of the cases, a single document suffices. Also, for 26% of cases, more than one candidate is plausible, (this is often due to hypernymy).
However, the questions of the WikiHop dataset are not in natural text form, which makes it different from real-world cases. Also, it is created by documents (from Web or Wikipedia) and a knowledge base (KB), as a result, these datasets are constrained by the schema of the KBs they use, and therefore the diversity of questions and answers is inherently limited [14]. also, these datasets have no information to explain the predicted answers [35].
Construction process: The process began with facts (subject, relation, object) from WikiData. Queries were created by taking the subject and relation, with the object being the answer. To ensure multi-hop reasoning, relevant supporting documents were identified by performing a breadth-first search on a bipartite graph connecting entities in WikiData to documents in WikiReading where those entities were mentioned. This process aimed to select documents where the information needed to answer the question was spread across multiple sources. Finally, a set of candidate answers, including the correct one and plausible distractors, were generated, creating a multi-choice question-answering task requiring inference across several documents. Also most models can be evaluated by the explainability by show the chain of reasoning process, but the dataset itself doesn't provide the supporting facts directly. Error analysis in WikiHop and MedHop often points to entity disambiguation across multiple documents or a lack of understanding of the complex relationships between entities as described in the supporting text. Also, analysis in MedHop can arise from the complex scientific terminology and relationships within the biomedical domain, requiring a deeper understanding than general-purpose models might possess.

Table 8. Dataset Sizes for the WikiHop and MedHop Datasets [34].

	Train	Dev	Test	Total
WikiHop	43738	5129	2451	51318
MEDHOP	1620	342	546	2508

Table 9. Type of the Questions in WikiHop [34].

Type of question	Percentage (%)
multi-step answer	45
Multiple plausible answers	15
Ambiguity due to hypernymy	11
Only single document required	9

HotpotQA [14] has become one of the most popular multi-hop MRC datasets in recent years. It has 113k Wikipedia-based question-answer pairs. To answer the questions of this dataset, it is required to collect and reason over information from multiple supporting documents. In contrast to WikiHop [34], the question and answer are in plain text form. The answers are variable-length spans, and there are 10 wiki passages per question. One of the distinguishing features of this database is that there are sentence-level supporting facts for each answer which is the sentences containing information that supports the answer. Thus, models should extract a span as the answer and also provide the supporting facts for the answers. The types of answers are shown in Table 10.
However, there are some drawbacks to this dataset, for example, QASC[33] discussed that since the questions of this dataset were authored in a similar way, due to their domain and task setup, they are easy to decompose, Also, as discussed in Inoue et al [36], the task of classifying sentence-level SFs is a binary classification task that is incapable of evaluating the reasoning and inference skills of the model. Besides, (Chen and Durrett, 2019; Min et al., 2019) revealed that many examples in this dataset do not require multi-hop reasoning to solve [35].
Construction process: The data collection process involved human annotators who were tasked with creating questions that necessitated combining information from at least two supporting paragraphs to arrive at the answer. Crucially, annotators also had to identify and annotate the specific supporting facts within the relevant paragraphs that justify the answer. This emphasis on providing explicit supporting evidence distinguishes HotpotQA and allows for models to be evaluated not only on answer accuracy but also on their ability to identify the necessary reasoning steps. Error analysis in HotpotQA often highlights misunderstanding comparative or superlative questions that require aggregating information across different parts of the text.

Table 10. Types of Answers in the HotpotQA Dataset [14].

Answer type	%	Example(s)
Person	30	King Edward , Rihanna
Group / Org	13	Cartoonito, Apalachee
Location	10	Fort Richardson, California
Date	9	10th or even 13th century
Number	8	79.92 million, 17
Artwork	8	Die schweigsame Frau
Yes/No	6	-
Adjective	4	Conservative
Event	1	Prix Benois de la Dance
Other	6	Cold War, Laban Movement
Noun		Analysis
Common noun	5	Comedy, both men and women

Table 11. Statistical Information of CMRC [37].

	Train	Dev	Test
Number of questions	10,321	3,351	4,895
Answer per question	1	3	3
Average passage token	452	469	472
Average answer tokens	17	9	9

CMRC [37] Provided a span-extraction MRC Dataset with more than 18,000 question which requires multi-sentence inference throughout the context. To build this dataset first, the Chinese portion of Wikipedia webpages are used as the context then, Traditional Chinese characters were converted into Simplified Chinese characters for normalization purpose. The questions in this dataset are completely annotated by human experts, which is different from previous datasets that rely on the automatic data generation. Some useful information of this dataset is shown in Table 11.

They also annotated a challenge as Table 12 in order to examine how well can reading comprehension models deal with the questions that need comprehensive reasoning over various clues in the context. Construction process: The data collection process involved selecting passages from diverse Chinese text sources, such as news, Wikipedia, and blogs. For each passage, human annotators formulated questions that could be answered by extracting a span of text directly from the passage. The annotators were instructed to create questions that varied in difficulty and required different levels of understanding of the text. Error analysis would center on the types of extractive errors made by models in Chinese. This could involve examining

cases where the predicted answer span is partially correct, completely wrong, too short or too long, or located in the wrong part of the passage.

Table 12. Statistical Information of the Challenge Test of CMRC [37].

	Challenge set
Number of questions	504
Answer per question	3
Average passage tokens	464
Average answer tokens	19

R^4C [36] is proposed to evaluate the internal reasoning of the reading comprehension system in HotpotQA [14]. As it's said before, HotpotQA requires identifying supporting facts (SFs), but since only a subset of SFs contributes to the necessary reasoning, thus, achieving high accuracy performance in the Support task cannot prove a model's reasoning ability. R^4C contains 4,588 questions and aims to evaluate a model's internal reasoning in a finer-grained manner than the Support task in HotPotQA. It requires giving not only answers but also derivations. A derivation is a semi-structured natural language form that is used to explain and justify the predicted answers. Each question is annotated with 3 reference derivations (i.e., 13.8k derivations). However, the small size of the dataset is a serious limitation to train and evaluate end-to-end systems [35]. Table 13 shows the statistics of the R4C corpus. "st." denotes the number of derivation steps. Each instance is annotated with 3 golden derivations.
Construction process: The R^4C dataset, designed by first selecting questions from the HotpotQA dataset[14]. Then, through a crowdsourcing framework, each question was annotated with three distinct reference derivations, which are explanations justifying the predicted answers as a sequence of reasoning steps represented as triples. This process resulted in a quality-assured dataset of approximately 4.6k questions, each paired with multiple human-generated derivations, enabling a more reliable assessment of a system's ability to not only answer correctly but also to provide the reasoning behind its answer. These derivations are explanations that justify the answers.

2WikiMultiHopQA [35] is a multi-hop dataset based on Wikipedia and WikiData , and consists of 192k multi-hop questions. An important point about this dataset is that it contains both structured and unstructured data. There are four types of questions in this dataset: 1) Comparison questions which are used to compare two or more entities, 2) Inference questions to infer relation r from the two relations r1 and r2, 3) Compositional questions to achieve the answer from an entity and two relations without any inference relations, 4) Bridge comparison which requires finding bridge entities. each sample contains some information, namely evidence that is used to explain the answer. Evidence is in from of triples, where each triple is structured data (subject entity, property, object entity) obtained from WikiData. You can see the number of examples in each type along with the average length of questions and answers of 2WikiMultiHopQA dataset in Table 14. This dataset like HotPotQA [14] uses the Answer, Support and Joint evaluation metrics.

Table 13. The Number of Derivations in R^4C [36].

Split	**QA#**	**#Derivations**			
		2st	3st	4st	Total
Train	2,379	4,944	1,553	640	7,137
Dev	2,209	4,424	1,599	604	6,627
Total	4,588	9,368	3,152	1,244	13,764

Construction process: The process is begun with selecting single-hop questions and answers from WikiQA dataset. Then crowed-workers looked for bridge entities in answers sentence for each question-answer pair. Using these bridge entities crowd workers formulate new multi-hop questions. Also, for each question the evidence information is provided which contains a reasoning path that can be used to evaluate the model explainibility. Error analysis in 2WikiMultiHop frequently highlights difficulties in understanding the relationships between entities mentioned across the two articles and performing the necessary logical inference to connect them and arrive at the answer.

Table 14. Question and Answer Lengths Across the frequency Questions [35].

Type of question	**Number of Examples**	**Avg. length of question**	**Avg. length of answer**
Comparison	57,989	11.98	1.58
Inference	7,478	8.41	3.15
Compositional	86,979	11.43	2.05
Bridge-comparison	40,160	17.01	2.01
Total	192,606	12.46	1.94

HybridQA [38] is a large-scale multi-hop QA dataset with heterogeneous data based on Wikipedia tables and their relevant texts and hyperlinks, then to answer the questions, tabular and textual information should be combined. The dataset consists of 70K question-answer pairs with 13K Wikipedia tables. To generate this database, some HITs (human intelligence tasks) have been defined. In each HIT, a table and its content have been presented to a crowd worker to create 6 questions and their corresponding answers. Tables have been limited to tables with 5 to 20 rows and 3 to 6 columns. For each hyperlink in the table, the first 12 sentences of the Wikipedia page have been retrieved. The types of multi-hop reasoning used in this dataset are as follows: 1) Table → Passage chain: it first finds the cells in the table, and then hop to their neighboring hyperlinked cells within the same row, finally extracts a text span from the passage of the hyperlinked cell as the answer. 2) Passage → Table chain: it first finds the related passage, which traces back to certain hyperlinked cells in the table, and then hop to a neighboring cell within the same row, finally extracts text span from that cell. 3) Passage → Table →Passage chain: it is similar to Type 2, but in the last step, it hops to a hyperlinked cell and extracts the answer from its linked passage. 4)Passage and Table jointly are used to identify a hyperlinked cell based on table operations and passage similarity and then extract the plain text from that cell as the answer. 5) involves two parallel reasoning chains, while the comparison is involved in the intermediate step to find the answer. 6)involve multiple reasoning chains, while superlative is involved in the intermediate step to obtain the correct answer. The percentage of each type of reasoning has been shown in Table 15.

Construction process: Its creation involved formulating questions that require reasoning over both structured data (tables, often from Wikipedia) and unstructured text (associated Wikipedia articles) to find the answer. 13K HITs (human intelligence task) on the Amazon Mechanical Turk platform is released, where each HIT presents the crowd-worker with one crawled Wikipedia table along with its hyperlinked passages. The workers are asked to write six questions and answers. Error analysis in HybridQA often highlights difficulties in misalignments between the structured and unstructured data or a lack of a unified framework for processing both modalities effectively.

Table 15. The Percentage of Each Type of Reasoning (Chen et al., 2020).

Type of reasoning	Percentage (%)
Table → Passage	23.4
Passage → Table	20.3
Passage → Table →Passage	35.1
Passage and Table	17.3
Two parallel reasoning chain	3.1
multiple reasoning chains	0.8

QASC [33] is a multi-hop dataset that contains 9,980 8-way multiple-choice questions from elementary and middle school level science, and ach question is produced by composing two facts from an existing text corpus. To generate this dataset, crowd workers have first received only the first fact fS. Then, they have freely chosen the second fact fL from other available facts. Finally, the questions have been created by compromising these two facts. Thus, the models require commonsense reasoning for the composition of these facts to find the answer. You can see the number of questions, and unique fS and fL in Table 14. In contrast to OpenbookQA that which the number of facts to answer a question is unclear, in QASC, it is clear that two facts are sufficient to answer a question. The number of questions, fS and fL has been shown in Table 16. Also, a comparison information among this dataset and previous multi-hop dataset has been shown in Table 17. As you can see QASC in addition to preparing supporting facts, annotated them too, also Decomposition is not evident.

Construction process: The data generation process involved starting with a set of declarative sentences expressing scientific facts. Then, questions were formulated that required synthesizing information from two or more of these factual sentences to arrive at the correct answer. The answers themselves are often compositions of phrases from the original sentences. This controlled generation process allows for targeted evaluation of a model's capacity to perform multi-sentence reasoning and information integration. Error analysis in QASC frequently points to a lack of understanding of the underlying scientific concepts or the logical relationships between the provided facts.

MuSeRC [39] is the first multi-hop MRC Russian dataset. The questions in Russian Multi-Sentence Reading Comprehension (MuSeRC) dataset rely on multiple sentences and commonsense knowledge. It contains more than 900 paragraphs and 12,805 sentences across 5 different domains, namely: (1) elementary school texts, (2) news, (3) fiction stories,

Table 16. The Number of the Questions, Unique fS and fL in QASC [33].

	Train	Dev	Test
Number of questions	8,134	926	920
Number of unique fS	722	103	103
Number of unique fL	6,157	753	762

Table 17. Compression Information [33].

Property	CompWebQ	HotpotQA	MultiRC	OpenBookQA	WikiHop	QASC
Supporting facts are available	N	Y	Y	N	N	Y
Supporting facts are annotated	N	Y	Y	N	N	Y
Decomposition is not evident	N	N	Y	Y	Y	Y

(4) fairy tales, and (5) summaries of TV series and books (Figure 8). In this dataset, the answers are not necessarily a span of the passage, and there may also be more than one answer to a question. Thus, the models should choose the best possible answer. Questions and answers have been created by crowd workers and, questions that can be answered with only one sentence have been removed to ensure that all questions are multi-hop. Although the number of the question in MuSeRC are less than MultiRC, but all of its questions are multi-hop (less than 60% question of MultiRC is multi-hop question), you can see some comparison information between MuSeRC and MultiRC in Table 18.

Construction process: The collected data is extracted from 5 different domains including (1) elementary school texts, (2) news, (3) fiction stories, (4) fairy tales, and (5) summaries of TV series and books. Then the crowd-workers were required to: (1) pose a natural language question to a given passage, (2) specify sentence numbers needed to obtain the answer, and (3) provide a set of both correct and incorrect answers. Error analysis in MuSeRC often highlights difficulties in failure to capture the implicit information, or the overall narrative arc conveyed by the multi-sentence context, leading to incorrect answers that require a holistic understanding.

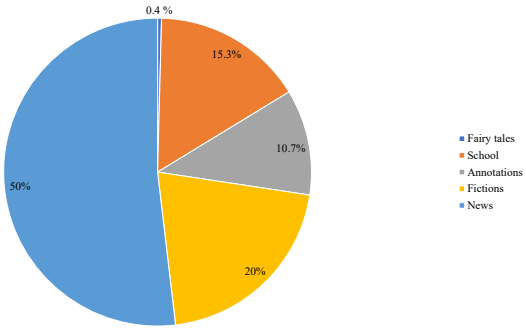


Figure 8. The Frequency of Domains [39].

Table 18. MuSeRC and MultiRC Comparison [39].

Property	MultiRC	MuSeRC
number of multi-hop questions	5,825	5,228
candidates / question	5.28	4.16
answers / question	2.31	1.86
sentence / passage	14.3	13.875
tokens / passage	258.9	203.9
tokens / question	10.9	7.61
tokens / answer	4.7	5.3
yes/no/true/false questions (%)	27.57%	1%

GCRC [40] proposed a reading Comprehension dataset on Gaokao Chinese (Chinese subject from the National College Entrance Examination of China), with more than 8,000 multi-choice questions and 5,000 passages. This dataset has focused on explainability of the dataset then there are three types of evidence of reasoning process: 1) sentence-level relevant (supporting facts), 2) error reason of an incorrect option, and 3) types of reasoning skills required for answering questions. Some useful statistical information of this dataset has been shown in Table 19.

Construction process: The data collection involved taking 10000 latest multiple-choice question and real and mock GaoKao Chinese from 5 websites. Then 12

graduate students annotate these questions individually and subsequently discuss and reach agreements if there is any discrepancy between two annotators. For each question, the specific sentences for the correct answer is also annotated which make the dataset appropriate to evaluate the explainability of the models. Error analysis would revolve around the reasons behind incorrect word predictions in the German cloze tasks. This might involve classifying errors based on grammatical categories (e.g., noun, verb, adjective), the type of linguistic knowledge required (e.g., morphology, syntax, semantics, discourse coherence), the distance of the context needed to infer the missing word, and the frequency or ambiguity of the correct word.

Table 19. Statistical Information of GCRC [40]

	Train	**Dev**	**Test**
Number of questions	6994	863	862
Number of questions with SF	2021	863	863
Number of questions with ER	2000	863	863
Number of questions with Ski	-	863	863

ComQA [41] is a large-scale Chinese dataset that contains more than 120k human-labeled questions. They focus on compositional QA which means the answers have been obtained from multiple but discontinuous segments in the documents and the answers are not limited to one span but can be a combination of several separate sentences. Besides, as in this dataset HTML pages are used as the context, even a table or an image can be selected as a part of the answer. To generate the ComQA dataset, a question-document pair is obtained first from a Chinese search engine called Sogou Search , the content of the retrieved page is then converted into a list of sentences, and the crowd workers find the answer to the question through the context. In Figure 10, you can see the frequency and type of questions in this dataset.
There are different kinds of questions in this dataset: 1) Simple: questions about a single property of an entity. 2)Compositional: questions that require answering more primitive questions and combining them. 2)Temporal: questions that require temporal reasoning for deriving the answer. 3)Comparison: questions that need some kind of comparison. 4)Telegraphic: short questions in an informal manner similar to keyword queries. 4) Answer tuple: Where an answer is a tuple of connected entities as opposed to a single entity. Table 20 shows Results of the manual analysis of 300 questions.

Construction process: The original resource curated by Fader et al. contains 763M questions. Questions in the crawl are grouped into 30M paraphrase clusters based on feedback from WikiAnswers users. The questions then were manually examined by crwoed-workers to remove unclear and non-factoid question to ended up with 13.7 k questions. Error analysis in ComQA frequently points to lack of sufficient context or the presence of subjective or opinion-based answers.

Table 20. Types of Questions in ComQA [41].

Property	**Percentage** (%)
Compositional questions	23.4
Temporal questions	20.3
Comparison questions	35.1
Telegraphic question	17.3
Answer tuple	3.1

JaQuAD [42] proposed a Japanese Question Answering Dataset for Machine Reading Comprehension (JaQuAD), consists of 39,696 extractive question-answer pairs on Japanese (Table 21). The dataset has been collected in three steps: (1) collecting contexts, (2) generating question-answer pairs on those contexts, and (3) validating collected questions and answers. For stages (2) and (3), human annotators were selected through a qualification test. Then, only the annotators who generated fluent questions and produced answers in a consistent format participated in dataset collection.
Human annotators generated multiple question-answer pairs while reading a context. A question had to be answerable using the content of the corresponding context. And an answer must be a span in the context. For the multi-hop question with all annotators generated answers according to the criteria provided. The reason types in this dataset is consist of multiple sentence reasoning, and Logical reasoning types, and the multi-sentences in this dataset is 20.2% which is more than SQuAD (13.6) makes this dataset suitable for multi-hop MRC.
However, most of the answers are short, and only 4.1% are longer than 8 tokens, therefore, JaQuAD has lower coverage for long answers compared to other datasets.
Construction process: The data collection process involved selecting passages from Japanese Wikipedia articles. For each passage, human annotators created questions that could be answered by extracting a

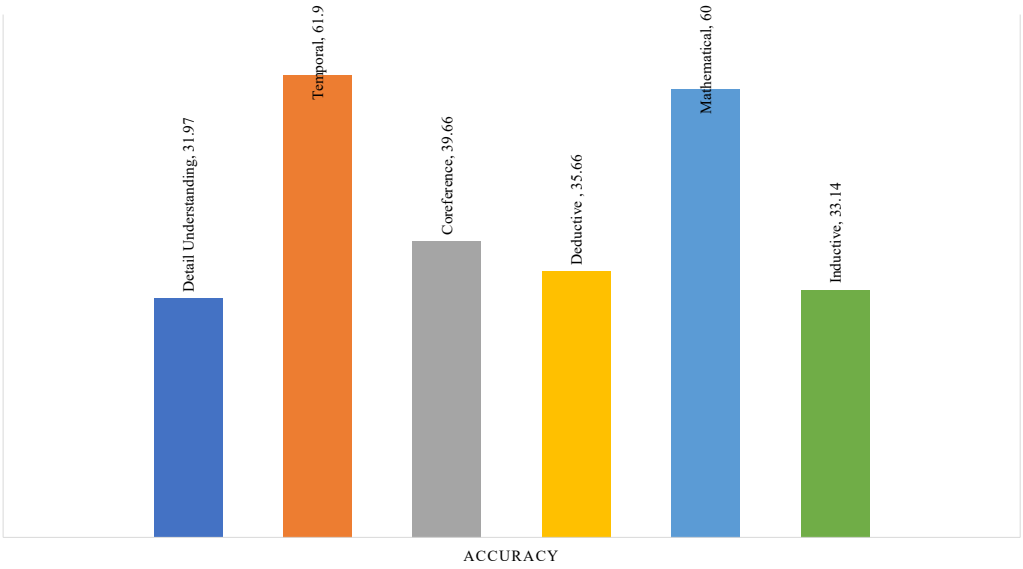


Figure 9. Results of BERT on GCRC.

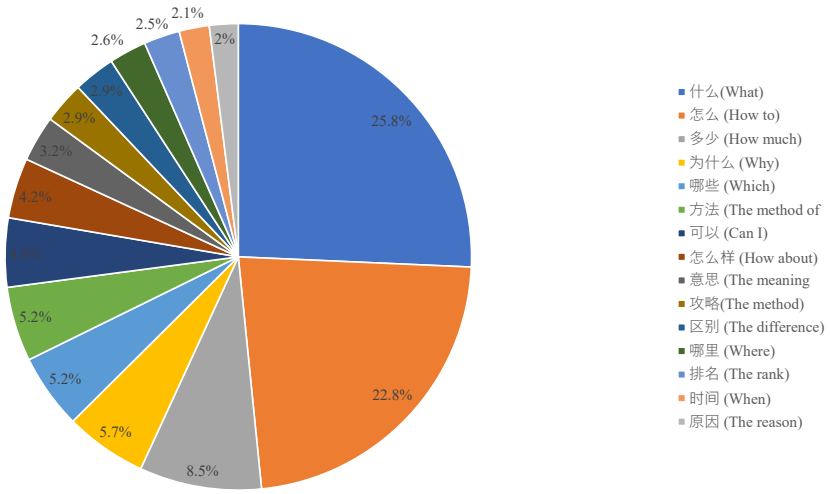


Figure 10. The Frequency and Type of Questions in ComQA [41].

span of text directly from the passage. The focus was on generating a diverse set of questions that covered various aspects of the passage content and required different levels of comprehension to identify the correct answer span within the Japanese text.

DSberQuAD [43] Proposed a Russian free-form question answering dataset named DSberQuAD with more than 2,000 questions. It is a modified version of SberQuAD. SberQuAD is a Russian dataset which more than 50k question and 9k paragraph. The building process of DSberQuAD can be divided into two steps: (1) Replace a question word in the question (e.g. who, when, why) with the short substring answer; (2) Form a sentence that follows grammatical agreement rules and change to the SVO word order. This dataset by choosing the most possible template and rank entities using the context and

Table 21. Statistical Information of JaQuAD [42].

	Total
Number of articles	800
Number of contexts	12,623
Number of questions	39,696

using pretrained BERT-base model to train systems components provided more accurate dataset.
Construction process: The data collection likely involved curating German documents from specific areas like legal texts, medical reports, or technical manuals. Human annotators then formulated questions based on these domain-specific passages, with the answers being spans of text directly present in the corresponding document. The error analysis would likely emphasize domain-specific challenges in German question answering. This would involve examining whether errors are related to the specialized vocabulary and terminology of the domain (e.g., legal, medical), the complex sentence structures often found in these domains, or a lack of background knowledge specific to the domain. Understanding if the errors are due to limitations in the pre-trained domain-specific BERT model or the question answering layer would be important.

ArabicaQA[44] introduced the large-scale and open-domain dataset for machine reading comprehension named ArabicaQA, to address the significant gap in Arabic natural language processing (NLP) resources. This dataset, consisting of 89,095 answerable and 3,701 unanswerable questions created by crowd workers to look similar to answerable ones, along with additional labels of open-domain questions marks a crucial advancement in Arabic NLP resources.
The main source of this dataset is articles from Arabic Wikipedia that were chosen randomly to prevent bias towards certain topics with a minimum of 20 words and a maximum of 500 words per paragraph. Then, the crowd workers generated question-answer pairs based on the paragraphs. These questions are identified for their depth and comprehensiveness, extending beyond mere length to provide a full understanding of the query.
Construction process: The data collection process involved selecting passages from various Arabic text sources, such as news articles, Wikipedia, and educational materials. Human annotators then created questions for each passage, where the answer to each question is a span of text that can be directly extracted from the passage. The dataset aims to facilitate the development of Arabic language understanding and question answering systems. Error analysis would focus on the complexity of Arabic morphology (e.g., prefixes, suffixes, infixes), the diglossic nature of the language (formal vs. colloquial), the right-to-left writing direction, and the nuances of Arabic syntax and semantics.
COMPLEXTEMPQA [45] introduced a novel dataset comprises over 100 million question-answer pairs, making it by far the largest dataset. It has important features as follows: 1) includes diverse categories of questions such as attribute-, comparison-, and counting-type questions, each relating to events, entities, or time periods, 2)Questions in COMPLEXTEMPQA require advanced reasoning skills, including event/entity matching, multi-hop question answering, across-time comparison, and temporal ordering 3)Each question is provided with detailed metadata, such as the time period it covers and a difficulty rating that classifies questions as “easy” or “hard” depending on the question type and expected familiarity of the entities involved. In Table 22 you can see some of reasoning types in COMPLEXTEMPQA dataset.

Construction process: The data collection involved creating question-answer pairs that necessitate understanding and comparing events and their timestamps. to do this every entry from single year pages from Wikipedia is extracted to gather information on significant events that happened over a span of 36 years (1987 to 2023). For instance, querying the item ”Presidents of the United States” within a specified year range. This often involved constructing questions that asked about the duration between events, the order of events, or events that occurred within a specific time frame. Error analysis would specifically target failures in reasoning about temporal information.

Table 22. The Reasoning Types in COMPLEXTEMPQA [45].

Name	Type
Comparison Event	25k
Comparison Entity	74k
Comparison Time	54k
Multi-hop	76k

CHRONICLINGAMERICAQA [46] is a question answering dataset with 487K questions built upon the vast collection of historical newspaper articles from the Chronicling America project. This dataset presents unique challenges due to the historical context, varied writing styles, and often lengthy and complex narratives found in old newspapers. It frequently includes multi-hop questions, requiring models to synthesize information from multiple sentences or paragraphs to arrive an answer. Table 23, Basis statistics of ChroniclingAmericaQA[46] shows some statistics information of this dataset.

Table 23. Basis Statistics of ChroniclingAmericaQA[46].

	Train	Dev	Test
#QA pairs	439,302	24,111	24,084
Average question length (words)	11.05	11.30	11.184
Average answer length (words)	2.01	1.978	1.981
Average question per newspaper page	54.80	4.07	4.08

Construction process: The data collection process of Chronicling America QA involves systematically scraping historical newspaper content from the Chronicling America website, a repository maintained by the Library of Congress. Using automated tools or APIs, the project retrieves newspaper titles, publication dates, metadata, and OCR text data spanning the 19th and early 20th centuries. After collection, the data undergoes cleaning and structuring to remove inconsistencies and errors, ensuring it's suitable for question answering tasks. This structured dataset then serves as the foundation for training and evaluating question-answering models focused on historical newspaper queries.
MRCEVAL [47] is a machine reading comprehension dataset designed to evaluate the robustness of models. It consists of adversarially generated questions, aiming to uncover vulnerabilities in models that might otherwise perform well on standard benchmarks. The questions are designed to be challenging, often requiring multi-hop reasoning to avoid being fooled by distractors or superficial cues. They also introduced a novel taxonomy that categorizes the key capabilities required for RC. Based on this taxonomy, MRCEval, an MRC benchmark with 2.1K high-quality multi-choice questions is constructed that leverages advanced Large Language Models (LLMs) as both sample generators and selection judges.

The needed RC skills for this dataset is summarized into three levels: context comprehension, external knowledge comprehension and reasoning. As, machines first understand the facts in the passage, then grasp the expression of the general information about the world by external knowledge, and the integration of facts and expertise for reasoning. Figure 11, MRCEval tasks division shows the MRCEval tasks division.

Construction process: The procedure begins with gathering a large collection of question-answer pairs, followed by careful curation to ensure diverse and high-quality reference answers. Next, human annotators create multiple valid answers per question, emphasizing varied phrasing and reasoning paths. These references are then standardized and formatted consistently. Finally, the dataset is split into evaluation subsets, with detailed documentation to facilitate usage in benchmarking generative models' performance against multiple plausible answers, addressing issues of subjectivity and enhancing fairness in evaluation.

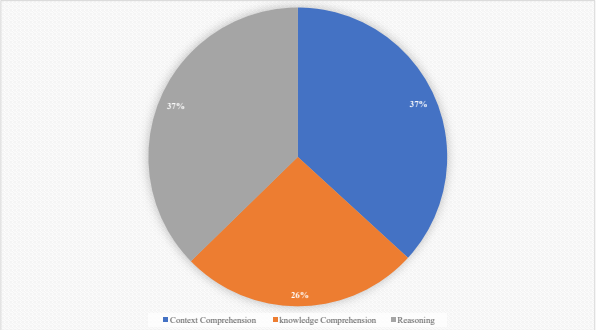


Figure 11. MRCEval Tasks Division.

5 Analysis

After reviewing each dataset in detail, some fine-grained comparisons will be presented in this section by preparing some figures and tables.

5.1 Frequency

First of all, the frequency usage of each dataset in recent multi-MRC models has been shown in Figure 11 (we used the same studies as Figure 2). As you can see, HotpotQA[14] and WikiHop [34] have received the most attention among the multi-hop datasets. There are several reasons for this:

(1) They are considered pioneers in multi-hop MRC because They were the first successful and serious attempt to cover multi-hop challenges, and they attract a great attention to the multi-hop MRC task.
(2) Although some datasets were introduced recently, since these two datasets have been used by many multi-hop models, they provide the good situation for comparing the performance of the models. Therefore, models would like to use these two datasets to compare their performance with other models, and this leads to more use of these two datasets.

We also prepare Figure 10 to investigate the growth trends of HotpotQA and WikiHop from 2018 to 2025 (note that both datasets have been introduced in 2018). As you can see, HotpotQA is more popular than WikiHop, there are some reasons for it:

(1) Questions in the WikiHop dataset are in triple form, while questions in the HotpotQA dataset are in the plain text form, which makes it more similar to real-world questions.
(2) Answers in HotpotQA are in the form of Span-extraction while answers in WikiHop are multiple-choice. As mentioned in Section 2, the Span-extraction task is more popular among models so the frequency usage of HotpotQA will be more.
(3) The HotpotQA dataset presents a new task called the Supporting fact task while WikiHop only focuses on the Answer task. So HotpotQA prepares a good situation to evaluate the ability of reasoning of the models.

Besides, as Figure 12 the number of usages of both datasets is maximum in 2019, which is due to the fact that the introduction of these two datasets in 2018 caused a lot of attention to the multi-hop MRC task, then in 2019, many multi-hop MRC models were proposed based on these two datasets. (It has been shown in Figure 2 too). In the following years, due to the introduction new multi-hop datasets, the usage of these two datasets got decreased but, both datasets are still the most popular datasets for multi-hop MRC models.

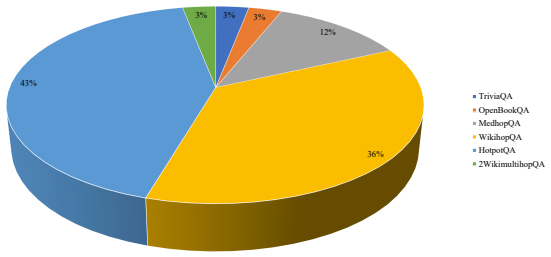


Figure 12. The Frequency of Each Dataset.

Table 24 contains some composition information of features of the reviewed multi-hop MRC datasets. In this table, 23 datasets are investigated, including general features like Domain, Language, Task, Number of questions, Release date, Average length of passages, and the number of passages per question. This information helps readers to compare existing datasets. Also, there is a feature about how many portions of each dataset are multi-hop. As it has been said before, we begin with some simple datasets that the questions can't be answered with a single sentence, although the sentences are within one passage and only a portion of these datasets include multi-sentence questions. These datasets are considered the first attempts to move from the single-hop datasets. (Note that some datasets are still used by recent multi-hop models, for example, DRNQA (X. Li et al., 2020) which has been published in 2020, has used TriviaQA [12]) With emerging datasets that especially have focused on the multi-hop challenge (2018), in addition to preparing more complex questions, the portions of the multi-hop questions have been increased in datasets (in most cases, all the questions are datasets).

5.2 Results

As the last review of multi-hop datasets in this section, the state-of-the-art result of existing multi-hop models on the reviewed multi-hop dataset will be shown. we considered the multi-hop MRC models from 2020 to 2025. As shown in Section 2, the two MRC tasks, Span-extraction and Multiple-choice, have received the most attention among datasets, so we will review them separately. It should be noted that due to the different models, the results are not expressed to compare the performance of datasets, but only to show the state-of-the-result on each database.

5.2.1 Span-extraction datasets:

Since the answers are in the span-extraction form the EM and F1 are used as the evaluation metrics. Table 25 shows the state-of-the-result of the span-extraction dataset alongside the models. There are some important points that can be discussed here. As it can be seen the results of the model on the Fullwiki setting of HotpotQA is relatively low than the Distractor setting, because the Fullwiki setting is more challenging, as it does not provide relevant paragraphs for the test set. Also, F1 score often yields better results than Exact Match which shows that EM is stricter than F1 because F1 measure overlap words and focus on the core information being covered but EM penalizes these variations. In addition, the size of the dataset (effective in the complexity and power of the model for training), the date of the dataset (effective to introduce new challenges), the number of multi-hop questions, the length of contacts (effective in reasoning process) which are mentioned in Table 24 should be considered.

5.2.2 Multiple-choice datasets:

Table 26 contains the stat-of-the-result of multi-hop models on multiple-choice datasets. Since the answer type of this dataset is multiple-choice then Accuracy is the evaluation metric on the test and development set.

Table 24. Statistical Information of Datasets.

	Dataset	Domain	Language	Task	Number of questions	Release date	Explainibility	Avg len. of passage	# of passages per question	Portion of multi-hop questions
1)	NewsQA (Trischler et al., 2017)	Open	English	Span-extraction	100K	2017	û	30.7 sentences	1 document	20.7
2)	Race(Lai et al., 2017)	Open	English	Multiple-choice	100K	2017	û	321.9 tokens	1 passage	25.8
3)	TriviaQA [12]	Open	English	Span-extraction	95K	2017	û	2,895 words	6 documents	40
4)	COMPLEXWEB (Talmor & Berant, 2018)	Open	English	Multiple-choice	43K	2018	û	13.18 tokens	—	100
5)	OpenBookQA (Mihaylov et al., 2018)	Science facts	English	Multiple-choice	6k	2018	ü	9.38 tokens	1.16 facts	100
6)	MultiRC ([7] et al., 2018)	Open	English	Cloze-style	9k	2018	û	263.1 tokens	1 paragraph	60
7)	WikiHop [34]	Open	English	Multiple-choice	51k	2018	û	100.4 tokens	13.7 documents	91
8)	MedHop [34]	Molecular Biology	English	Multiple-choice	2k	2018	û	253.9 tokens	36.4 documents	100
9)	HotpotQA [14]	Open	English	Span-extraction	113K	2018	ü	—	10 paragraphs	100
10)	CMRC	Open	Chinese	Multiple-choice	18,000	2019	û	452 tokens	—	—
11)	R4C[36]	Open	English	Multiple-choice	4,588	2020	ü	—	10 paragraphs	100
12)	2WikiMultiHopQA [35]	Open	English	Span-extraction	192K	2020	ü	—	10 paragraphs	100
13)	HybridQA (W. Chen et al., 2020)	Open	English	Free-form	70K	2020	û	15.7 rows, 4.4 columns	1 table	100
14)	QASC [33]	Science	English	Cloze-style	9K	2020	û	—	2 facts	100
15)	MuSeRC [39]	Open	Russian	Multiple-choice	5K	2020	û	1.5k characters	1 passage	100
16)	GCRC	Open	Chinese	Span-extraction	6K	2021	ü			
17)	ComQA [41]	Open	Chinese	Span-extraction	120K	2021	û	581 words	1 document	100
18)	JaQuAD	Open	Japanese	Span-extraction	39K	2022	û	266 words	—	20
19)	DSberQuAD	Open	Russia	Free-form	2K	2023	û	101 words	—	—
20)	ArabicaQA	Open	Arabia	Span-extraction	90K	2024	û	500 words	1 paragraph	100
21)	COMPLEXTEI	Open	Persian	Span-extraction	100,000K	2024	û	—	—	100
22)	CHRONICLING	Historical News Articles	English	Span-extraction	485K	2024	û	250 words	3 paragraphs	28
23)	MRCEval	Open	English	Multiple-choice	2.1K	2025	û	—	—	37

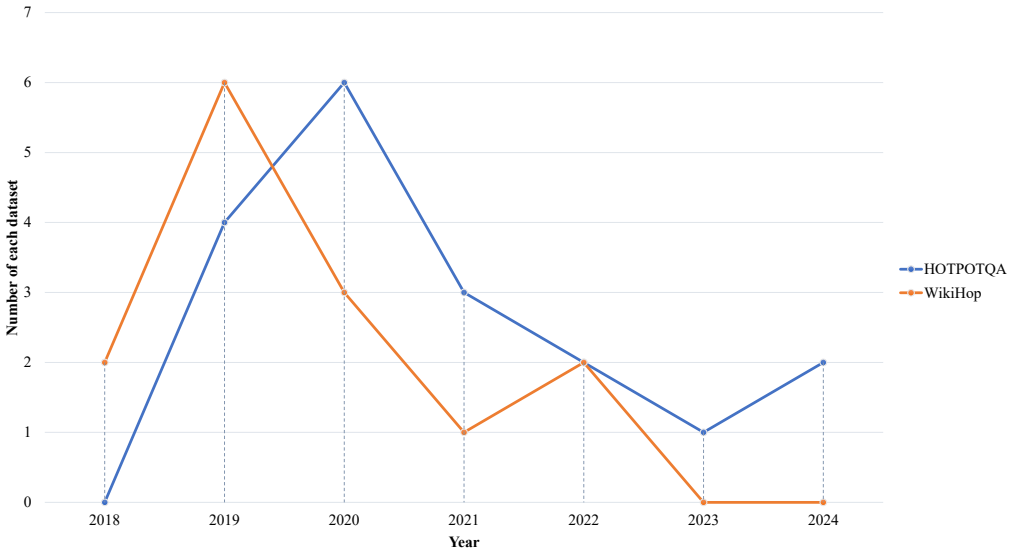


Figure 13. The HotpotQA and WikiHop growth Trend.

Table 25. The State-of-the-result on Span-extraction Multi--hop Datasets.

Dataset	Model	EM	F1
HotpotQA-distractor [14]	GPT-4 (P. Panda et al., 2024)	69.3	82.2
HotPotQA-fullwiki [14]	HGN (Fang et al., 2020)	57.85	69.93
2WikiMultiHopQA [35]	LM(Li.Y. et al. , 2024)	53.4	67.00
TriviaQa [12]	DRNQA (X. Li et al., 2020)	59.73	62.21
MuScRC [39]	MuScRC [39]	25.6	65.6
HybridQA [38]	HProPro(Shi. Qi et al. ,2024)	46.8	54.4
CHRONQA [46]	Gemma-2-2B (Team G. et al., 2024)	16.54	—

Table 26. The State-of-the-result on Multiple-choice Multi--hop Datasets.

Dataset	Model	Accuracy	
		Test set	Dev set
WikiHop [34]	ClueReader (Peng et al., 2023)	72.0	66.5
MedHop [34]	ClueReader (Peng et al., 2023)	46.00	48.2
OpenBookQA [32]	PathNet(Kundu et al., 2018)	53.4	55.0
QASC [33]	Flan-T5 (Mehrafarin et al., 2024)	77.0	—

5.2.3 Reasoning Compare

Because of the importance of the reasoning skill in multi-hop datasets in this section the new datasets like QASC, ComplexTempQA, and the most popular multi-hop datasets (HotpotQA) are investigated in more details. All of these datasets are designed to evaluate multi-hop reasoning in question answering, but they differ significantly in their focus, complexity, data sources, and question types (Table 27). QASC It emphasizes the ability to combine multiple facts from a corpus to answer a question. The decomposition into these facts is often not evident from the question itself.

Complexity: While multi-hop, the reasoning often involves relatively straightforward logical composition of scientific facts. ComplexTempQA Complex questions requiring temporal reasoning, including across-time comparison, temporal aggregation, and multi-hop reasoning involving temporal event ordering and entity recognition.

Complexity: High, due to the need to understand and reason about temporal relationships and spans. Explainability: The detailed metadata, including time scopes, can aid in understanding the model's reason-

Table 27. Reasoning Compare.

Dataset	Reasoning Task	Reasoning Method	Reasoning Complexity
		Test set	Dev set
HotpotQA	Answering questions requiring reasoning over multiple supporting documents (often 2).	Finding and combining information across multiple text paragraphs.	Diverse strategies including bridging entities, intersection, and comparison. Requires identifying relevant facts and performing multi-hop inference.
WikiHop	Multi-choice question answering where the answer requires reasoning across multiple documents.	Inference on a graph where nodes are entities and edges represent relations within and across documents (often using GCNs).	Requires identifying entities and relations across documents and performing multi-step reasoning to find the correct answer among candidates.
Com.TempQA	Answering temporal questions that require reasoning about events, entities, and time periods.	Across-time comparison, temporal aggregation, and multi-hop reasoning involving temporal event ordering and entity recognition. 1	High complexity, demanding understanding of temporal relations, event sequences, and comparisons over time, often spanning decades.
QASC	Question answering based on a corpus of scientific facts.	Combining multiple facts to answer a question that is not directly stated in a single fact.	Requires identifying relevant scientific facts and performing logical reasoning or inference to synthesize the answer. Complexity varies depending on the number of facts to combine and the type of reasoning (e.g., conjunction, disjunction).

ing over time.
HotpotQA ocus: Diverse, explainable multi-hop question answering. Questions are designed to require reasoning over multiple supporting documents (typically two Wikipedia article introductions).
Complexity: Moderate to high, involving various reasoning strategies like bridging entities, intersections, and comparisons. Supporting Facts: Annotations include the two gold paragraphs and the specific sentences within them that support them.

6 OPEN ISSUES

Datasets with the new challenges are a good motivation to improve current models. So, to reduce the gap between existing multi-hop MRC models and real-world cases, creating the datasets with more challenges is vital. Even the most popular multi-hop datasets still have shortcomings, then if models only focus on the existing dataset, the gap between them and real-world cannot be properly reduced, so it is important to focus on building datasets. The lower number of free-form answering and cloze-style multi-hop MRC datasets is primarily due to the increased complexity and cost of annotation and evaluation, which has historically led to a greater focus on other types of MRC tasks. The lack of free-form and close-style datasets has made it impossible to present models for these two tasks, while both tasks have a good potential to use in form of multi-hop MRC. Among all the types of MRC tasks described in Section 2, the free-form task is the most similar task to the real-world case but due to its complexity, there is no proper dataset for it. Since this type of answer is more similar to real-world scenarios, focusing on

presenting free-form datasets to encourage models can improve the application of MRC systems. Future research in question answering should explore less emphasized styles such as free-form and cloze-style QA, focusing on bridging the gap between extractive and generative QA through hybrid models that can flexibly handle both styles. This includes advancing multi-hop reasoning capabilities for cloze tasks, leveraging multi-task learning for cross-style generalization, and developing adaptive systems that can personalize responses based on user feedback. Additionally, improved evaluation metrics tailored to the nuances of free-form and cloze QA, along with data-efficient methods like few-shot and self-supervised learning, will be vital for expanding QA performance across diverse domains and user contexts Large Language Models (LLMs) such as ChatGPT and PaLM have significantly impacted the field of multi-hop question answering (QA) datasets. The emergence of LLMs, with their enhanced abilities in understanding context, performing complex reasoning, and generating coherent text, has led to both advancements and new challenges in the context of these datasets. One key impact is the surprisingly strong performance of LLMs on existing multi-hop QA datasets, often achieving state-of-the-art results with minimal or no fine-tuning. Then existing multi-hop QA datasets is not sufficient to evaluate the true reasoning abilities of LLMs. Moreover, LLMs are also being used as tools in the creation of multi-hop QA datasets. Their ability to generate complex questions and plausible distractor answers can aid in the development of more challenging and diverse benchmarks. In conclusion, LLMs have demonstrated impressive capabilities on multi-hop QA datasets, pushing the boundaries of what's achievable in this task. However, their performance has also highlighted the shortcomings of existing benchmarks and driven the community towards developing more robust and insightful evaluation methods that can truly assess the multi-step reasoning abilities of these powerful models. The interaction between LLMs and multi-hop QA datasets is an ongoing and evolving process, with each influencing the development of the other.

7 Conclusions

In this study, a comprehensive survey was presented with a focus on the multi-hop MRC dataset and evaluation metrics as two vital and important aspects of multi-hop MRC. In the beginning, the definition of the multi-hop MRC problem has been presented then the evaluation metrics have been investigated in detail. In following, 23 multi-hop datasets from 2017 to 2025 had been reviewed in detail with a strong focus on the multi-hop aspects of them, then, some comprehensive analyses were presented for a fine-grain comparison of the different features of datasets. Finally, open issues in this field had been discussed.

References

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics*, pages 2383–2392, 2016. doi:10.18653/v1/D16-1264.

[2] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701, 2015. doi:10.48550/arXiv.1506.03340.

[3] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the {CNN}/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany, August 2016. Association for Computational Linguistics. doi:10.18653/v1/P16-1223.

[4] Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017. doi:10.48550/arXiv.1707.03904.

[5] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602, Vancouver, Canada, 2017. Association for Computational Linguistics. doi:10.18653/v1/P17-1055.

[6] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. doi:10.48550/arXiv.1709.04696.

[7] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. Question answering via integer programming over semi-structured knowledge. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*,

volume 2016-January, pages 2232–2238, 2016. doi:10.48550/arXiv.1604.06076.

[8] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics, 2017. doi:10.18653/v1/D17-1215.

[9] Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735. Association for Computational Linguistics, 2018. doi:10.18653/v1/P18-1160.pdf.

[10] Feng Gao, Jiancheng Ni, Peng Gao, Zili Zhou, Yan-Yan Li, and Hamido Fujita. Heterogeneous graph attention network for multi-hop machine reading comprehension. *CoRR*, 2021. URL https://arxiv.org/abs/2101.11954.

[11] Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. Evidence integration for multi-hop reading comprehension with graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2020. doi:10.1109/TKDE.2020.2982894.

[12] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017. doi:10.18653/v1/P17-1147.

[13] Tomáš Kočisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi:10.1162/tacl_a_00023.

[14] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380, 2018. doi:10.18653/v1/D18-1259.

[15] Yichen Jiang and Mohit Bansal. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4474–4484, 2019. doi:10.18653/v1/D19-1455.

[16] Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698, 2019. doi:10.3390/app9183698.

[17] Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. A survey on machine reading comprehension systems. *Natural Language Engineering*, pages 1–50, 2020. doi:10.48550/arXiv.2001.01582.

[18] Mokanarangan Thayaparan, Marco Valentino, and André Freitas. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*, 2020. doi:10.48550/arXiv.2010.00389.

[19] Zhuosheng Zhang, Hai Zhao, and Rui Wang. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*, 2020. doi:10.48550/arXiv.2005.06249.

[20] Yang Bai and Daisy Zhe Wang. More than reading comprehension: A survey on datasets and metrics of textual question answering. *arXiv preprint arXiv:2109.12264*, 2021. URL https://arxiv.org/abs/2109.12264.

[21] Danqi Chen. *Neural reading comprehension and beyond*. Stanford University, 2018.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. doi:10.3115/1073083.1073135.

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[24] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*, 2005. doi:10.3115/1220575.1220629.

[25] Qi Liu, Matt J. Kusner, and Phil Blunsom. A survey on contextual embeddings. *DeepAI*, 2020. doi:10.48550/arXiv.2003.05112.

[26] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016. doi:10.48550/arXiv.1611.09830.

[27] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale read-

ing comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics, 2017. doi:10.18653/v1/D17-1082.

[28] Alon Talmor and Jonathan Berant. Repartitioning of the complexwebquestions dataset. *arXiv preprint arXiv:1807.09623*, 2018. doi:10.48550/arXiv.1807.09623.

[29] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, 2014. doi:10.3115/v1/D14-1159.

[30] Wen tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, 2016. doi:10.18653/v1/P16-2033.pdf.

[31] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, 2008. doi:10.1145/1376616.1376746.

[32] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, 2018. doi:10.18653/v1/D18-1260.

[33] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090, 2020. doi:10.48550/arXiv.1910.11473.

[34] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. doi:10.48550/arXiv.1710.06481.

[35] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, 2020. doi:10.18653/v1/2020.coling-main.580.

[36] Naoya Inoue. R4c: A benchmark for evaluating rc systems to get the right answer for the right reason. *Journal of Natural Language Processing*, 27, 2020. ISSN 1340-7619. doi:10.5715/jnlp.27.665.

[37] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for chinese machine reading comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 5883–5889, 2019. doi:10.18653/v1/D19-1600.

[38] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Proceedings of Findings of EMNLP 2020*, pages 1026–1036, 2020. doi:10.18653/v1/2020.findings-emnlp.91.

[39] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. Read and reason with muserc and rucos: Datasets for machine reading comprehension for russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6481–6497, 2020. doi:10.18653/v1/2020.coling-main.570.

[40] Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. Gcrc: A new challenging mrc dataset from gaokao chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, 2021. doi:10.18653/v1/2021.findings-acl.113.

[41] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. *arXiv preprint arXiv:1809.09528*, 2018. doi:10.18653/v1/N19-1027.

[42] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. Jaquad: Japanese question answering dataset for machine reading comprehension. *arXiv preprint arXiv:2202.01764*, 2022. doi:10.48550/arXiv.2202.01764.

[43] Raushan Turganbay, Viacheslav Surkov, Dmitrii Evseev, and Mikhail Drobyshevskiy. Generative question answering systems over knowledge graphs and text. pages 1112–1126, 2023. doi:10.28995/2075-7182-2023-22-1112-1126.

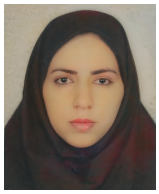
[44] Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Ja-

towt. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059, 2024. doi:10.48550/arXiv.2403.17848.

[45] Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. Complextempqa: A large-scale dataset for complex temporal question answering. *arXiv preprint arXiv:2406.04866*, 2024. doi:10.48550/arXiv.2406.04866.

[46] Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2038–2048, 2024. doi:10.1145/3626772.3657891.

[47] Shengkun Ma, Hao Peng, Lei Hou, and Juanzi Li. Mrceval: A comprehensive, challenging and accessible machine reading comprehension benchmark. *arXiv preprint arXiv:2503.07144*, 2025. doi:10.48550/arXiv.2503.07144.



Azade Mohammadi Dr. Azadeh Mohammadi received her B.Sc. and M.Sc. degrees in Software Engineering from Razi University, Kermanshah, and her Ph.D. in Software Engineering from the University of Isfahan. Her research interests include machine learning, natural language processing (NLP), data mining, and human-computer interaction (HCI). She has published research papers in reputable journals and conferences and has several years of teaching experience at universities such as the University of Isfahan, Razi University, and Kermanshah University of Technology. In addition to her academic background, Dr. Mohammadi has professional experience as an IT specialist, managing and supporting information systems in organizational settings.

Reza Ramezani Dr. Reza Ramezani received his B.Sc. in Computer Engineering from Shiraz University of Technology in 2010 and continued his M.Sc. in the same field at Isfahan University of Technology, completing it in 2012. He then joined the Department of Computer and Automation at UCM Madrid in 2014 and earned his Ph.D. in Software Engineering from Ferdowsi University of Mashhad in fall 2017. Currently, he serves as an Associate professor in the Software Engineering group at the University of Isfahan. His research focuses on natural language processing, data mining, and the semantic web. Over the past decade, he has taught undergraduate to Ph.D. courses, including advanced programming, object-oriented development, and data mining