UnAnswGen: A Systematic Approach for Generating Unanswerable Questions in Machine Reading Comprehension

Hadiseh Moradisani*
School of Engineering, University of Guelph
Guelph, Ontario, Canada
hmoradis@uoguelph.ca

Fattane Zarrinkalam
School of Engineering, University of Guelph
Guelph, Ontario, Canada
fzarrink@uoguelph.ca

Julien Serbanescu
School of Engineering, University of Guelph
Guelph, Ontario, Canada
serbanej@uoguelph.ca

Zeinab Noorian
Information Technology Management, Toronto Metropolitan University
Toronto, Ontario, Canada
znoorian@torontomu.ca

Abstract

This paper introduces a configurable software workflow to automatically generate and publicly share a dataset of multi-labeled unanswerable questions for Machine Reading Comprehension (MRC). Unlike existing datasets like SQuAD2.0, which do not account for the reasons behind question unanswerability, our method fills a critical gap by systematically transforming answerable questions into their unanswerable counterparts across various linguistic dimensions including entity swap, number swap, negation, antonym, mutual exclusion, and no information. These candidate unanswerable questions are evaluated using advanced MRC models to ensure their context-based unanswerability, with the final selection based on a majority consensus mechanism. Our approach addresses the scarcity of multi-labeled datasets like SQuAD2-CR, enabling comprehensive evaluation of MRC systems' ability to handle unanswerable queries and facilitating the exploration of solutions such as query reformulation. The resulting UnAnswGen dataset and associated software workflow are made publicly available to advance research in machine reading comprehension, offering researchers a standardized toolset for evaluating and enhancing MRC systems' robustness and performance.

CCS Concepts

• **Machine Learning and NLP for IR → Question answering**.

Keywords

Unanswerable Question, Machine Reading Comprehension, SQuAD2.0 dataset, Question Answering System

*Corresponding author: Hadiseh Moradisani.

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698413

ACM Reference Format:
Hadiseh Moradisani, Fattane Zarrinkalam, Julien Serbanescu, and Zeinab Noorian. 2024. UnAnswGen: A Systematic Approach for Generating Unanswerable Questions in Machine Reading Comprehension. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3673791.3698413

1 Introduction

Understanding text and responding to questions are fundamental for natural language processing [3, 21, 26, 36]. The creation of several large-scale datasets such as SQuAD1.0 [30] and MS MARCO [25] has driven notable advancements in Machine Reading Comprehension (MRC) tasks. However, a common assumption in many current methods is that a correct answer always exists within the context passage. As a result, these methods typically focus on selecting the most plausible text span based on the question, without confirming the presence of an answer [19, 29, 32]. Ideally, systems should avoid responding rather than making uncertain guesses, demonstrating their language comprehension abilities. Recently, the advent of datasets containing unanswerable questions, such as SQuAD2.0 [29], has drawn considerable attention from researchers focused on the challenge of unanswerability. A brief summary of these datasets is included in Table 1. Based on these datasets, several innovative methods have been proposed to address the unanswerability issue ranging from probability prediction and extraction [4, 12, 15, 18, 32] to leveraging large language models [1, 13, 14, 37, 38]. However, these studies typically categorize questions as either answerable or unanswerable, leaving a gap in the underlying linguistic causes behind their unanswerability. As MRC systems advance to meet the complexity of real-world information needs, there is an increasing demand to not only detect unanswerable questions but also to understand and diagnose the underlying factors that contribute to their lack of answerability [19]. A significant challenge in identifying the causes of unanswerability within the MRC domain is the limited availability of multi-labeled datasets. SQuAD2-CR is one of the few datasets that provide such data, as noted in Table 1. SQuAD2-CR [17] enriches the SQuAD2.0 by labeling its unanswerable questions with their specific reasons for

Table 1: Datasets used in the state-of-the-art MRC models including answerable and unanswerable questions (NA indicates Not Available).

Name	Data Source	Generation method	# of (Answerable - Unanswerable)	Unanswerability Causes	Citations	year
NewsQA [34]	CNN	Crowdsourcing	(102,146 - 5,528)	NA	[1, 15, 38]	2017
DuoRC [31]	Wikipedia,IMDB	Crowdsourcing	(58,752 - 10,772)	NA	[24]	2018
SQuAD-T [32]	Wikipedia	Crowdsourcing	(57,024 - 29,806)	NA	[32]	2018
SQuAD2.0 [29]	Wikipedia	Crowdsourcing	(86,821 - 43,498)	NA	[1, 7, 8, 12, 13, 19, 20, 22, 27, 33, 38, 39]	2018
UNANSQ [39]	Wikipedia	automated	(0 - 69,090)	NA	[8]	2019
CRQDA [20]	Wikipedia	automated	(0 - 124,085)	NA	[8]	2020
SQuAD2-CR [17]	Wikipedia	Crowdsourcing	(86,821 - 43,498)	6	[19]	2020
Dureader [11]	Chinese search engine	user-logs	(258,475 - 13,099)	NA	[27]	2022
IDK-MRC [28]	Wikipedia	automated and crowdsourcing	(5,042 - 4,290)	NA	-	2022
Lightweight Dataset [8]	Wikipedia	automated	(0 - 81,804)	2	-	2023

their unanswerability, including *Entity Swap*, *Number Swap*, *Negation*, *Antonym*, *Mutual Exclusion*, and *No Information*. Utilizing this dataset, research in [19] is one of the few studies that employ a multi-class classification approach to attribute unanswerable questions by identifying the particular causes for their lack of answerability. However, the SQuAD2-CR dataset faces challenges due to the scarcity of data in certain categories of unanswerability. For instance, it contains only $3,350$ unanswerable questions labeled with *No Information*. Moreover, as this dataset was created using a crowdsourcing approach, expanding it is non-trivial and incurs substantial costs. Recent studies like UNANSQ [39], CRQDA [20] and AGent framework [33] have attempted to automatically generate unanswerable questions. However, these approaches do not provide labels or identify the types of unanswerability for the questions they generate, highlighting a gap in the automated generation and categorization of unanswerable questions within current research.

To address this challenge, we propose an automated method to expand these datasets with a broader spectrum of unanswerable questions across diverse categories. This expansion facilitates the improvement of the evaluation of systems' capabilities in detecting unanswerable questions and enables the exploration of various causes of unanswerability. The only study similar to our work is the recent research in [8], which focuses on the automatic generation of multi-label unanswerable questions through antonym and entity augmentations alone. In contrast, our method encompasses a wider range of reasons for unanswerability, including entity swap, number swap, negation, antonym, mutual exclusion, and no information, as outlined in [29]. Additionally, our approach is designed to generate unanswerable versions of answerable questions within their original context. This capability not only enhances the detection of unanswerable questions but also empowers researchers to develop solutions, such as query reformulation, aimed at transforming unanswerable questions into answerable ones.

To develop a multi-label MRC dataset with unanswerable questions, we propose a configurable software workflow that takes as input a set of answerable questions along with their associated context passage and correct answers, e.g., SQuAD2.0. The output is a comprehensive dataset that includes a list of unanswerable questions for each of the answerable questions in the input along with reasons for their unanswerability. This is accomplished in two main steps. First, a host of state-of-the-art unsupervised techniques are implemented to systematically generate a large pool of candidate unanswerable questions across six different categories: *Entity Swap*, *Number Swap*, *Negation*, *Antonym*, *Mutual Exclusion*, and *No Information* for each input question. Second, the generated candidate unanswerable questions are evaluated using existing state-of-the-art MRC models to determine their unanswerability based on the associated context passage. Those questions that are identified as unanswerable by the majority vote of the models are selected for inclusion in the output dataset. The specific linguistic approach used to generate each candidate unanswerable question is recorded, serving as the label for that question.

Using this configurable software workflow, we have developed an MRC dataset featuring multi-labeled unanswerable questions, named the *UnAnswGen* dataset. We have made the code, the executable workflow, and the generated dataset publicly available and can be accessed via the link[1].

The advantages of our work are twofold: (1) Our implementation of the proposed software workflow allows community members to automatically generate new multi-labeled MRC datasets with unanswerable questions for any input MRC dataset; and (2) We provide an "out of the box" MRC dataset based on SQuAD2.0, which includes six types of unanswerable questions. This dataset is immediately available for use, saving researchers the time and effort required to generate and validate their own unanswerable questions.

2 Proposed Workflow

In this section, we outline our proposed workflow for automatically generating unanswerable questions labeled by various classes of unanswerability for an MRC dataset. We then describe how we used this process to create the *UnAnswGen* dataset, an augmented version of SQuAD2.0 that includes multi-labeled unanswerable questions. Figure 1 presents the overview of our proposed workflow.

The input of this workflow consists of a standard MRC dataset represented by a set of triples $M = \{(q_i, p_i, a_i) | q_i \in Q_a, p_i \in P, a_i \in A\}$, where Q_a represents a set of *answerable questions*, P denotes their associated context passage, and A denotes their corresponding golden truth answer. The output of the workflow is a set of unanswerable questions for each answerable question $q_i \in Q_a$, denoted by U_{q_i}. Let L represent the set of possible classes of unanswerability, each unanswerable question $q'_i \in U_{q_i}$ is labeled with a class of unanswerability $l_i \in L$ which indicates the reason why q'_i is unanswerable compared to its answerable version q_i, based on its associated context passage p. Our proposed workflow includes two components: (1) Unanswerable Question Generation and (2)

[1]https://github.com/Julien-ser/UnAnswGen

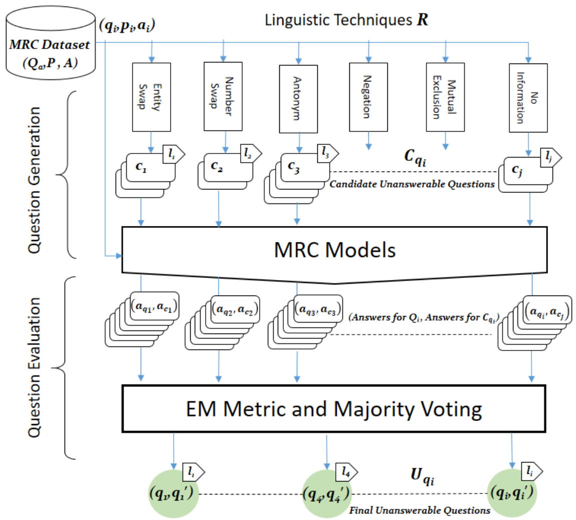

Figure 1: Overview of the UnAnswGen framework.

Unanswerable Question Evaluation, which are elaborated in the subsequent sections.

2.1 Unanswerable Question Generation

The purpose of this component is to generate a set of Candidate Unanswerable Questions (C_{q_i}) for each answerable question in an MRC dataset, based on potential reasons for unanswerability. Drawing inspiration from the taxonomy of unanswerability in MRC systems as outlined in [29], we systematically apply a variety of techniques to transform answerable questions in the input MRC dataset into their unanswerable counterparts, addressing various linguistic dimensions such as entity swap, number swap, negation, antonym, mutual exclusion, and no information.

Formally, in this step, for a given answerable question $q_i \in Q_a$ and its associated context passage p_i in dataset M, a list of candidate unanswerable questions C_{q_i} is generated using a set of linguistic techniques R designed to produce potential unanswerable versions based on each reason of unanswerability:

$$C_{q_i} = \bigcup_{r \in R} r(q_i) \tag{1}$$

Here, C_{q_i} consists of pairs (c_j, l_j) where $l_j \in L$ represents the set of possible classes of unanswerability. Thus, given a set of answerable questions Q_a and their corresponding passages P, the output includes a list of triples $\{(q_i, p_i, C_{q_i}) | q_i \in Q_a, p_i \in P\}$ where each triple consists of an answerable question, its associated context passage, and its candidate unanswerable questions. We have implemented and integrated a comprehensive set of linguistic techniques represented by R for generating unanswerable questions, tailored to each category of unanswerability. Below, we briefly outline these techniques corresponding to each class of unanswerability.

Entity Swap. According to the definition of Entity Swap, a question can become unanswerable by substituting one entity with another [29]. Thus, in our implemented method to generate the unanswerable version of each answerable question for a given context passage by entity swap approach, we apply two main steps: (1) *Extraction*: we utilize Spacy [2] to first extract all entities and their types (e.g., noun, verb, pronoun) from both q_i and p_i. (2) *Replacement*: For each entity in the input answerable question, we replace it with another entity of the same type extracted from its associated context p_i. For example, the input answerable question *In what city and state did Beyoncé grow up?* can be transformed into the candidate unanswerable question *In what city and state did Mathew Knowles grow up?*. In this example, the corresponding context passage of the input question mentions only Beyoncé's birth and upbringing, identifying Mathew Knowles merely as her father without providing details about his birth. Therefore, the second question becomes unanswerable based on the given context. By considering that the replacement involves entity types, we maintain readability and grammatical correctness, preserve the same answer type as the original question, and ensure relevance to the corresponding context.

Number Swap. Number Swap involves modifying a question to potentially render it unanswerable by replacing numbers with other numbers from the context. Similar to the Entity Swap method, we utilize an entity extraction method to generate the unanswerable version of each answerable question for a given passage context. This process involves two main steps: (1) *Extraction*: We extract all numerical entities and their types (e.g., cardinal numbers, ordinal numbers) from both the input question and its context using Spacy. (2) *Replacement*: For each numerical entity in the input question, we replace it with another numerical entity of the same type extracted from the context passage. For instance, based on this approach, a potential unanswerable version of the input answerable question *Time magazine named her one of the most 100 influential people of the century?* could be *Time magazine named her one of the most 19 influential people of the century?*. Here, the numerical entity "100" in the input question is replaced with "19", which is extracted as a numerical entity of the same type from the corresponding passage context. By ensuring that the replacement considers numerical entity types, we maintain grammatical correctness, preserve the structure of the original question, and ensure relevance to the corresponding context.

Antonym. Antonym modification involves altering a question to potentially render it unanswerable by replacing words with their antonyms. The steps are as follows: (1) *Extraction*: Identify words in the input question that have antonyms using WordNet [6]. (2) *Replacement*: Replace each identified word with its antonym, ensuring the modified question remains contextually appropriate. For instance, the input answerable question: *When did Beyoncé leave Destiny's Child and become a solo singer?* can be modified to *When did Beyoncé enter Destiny's Child and become a solo singer?*. Here, the word "leave" in the input question is replaced with "enter", its antonym, maintaining the grammatical structure and relevance to the context.

Negation. Negation Modification involves altering questions to potentially render them unanswerable by inserting or removing negation words such as "not" and "never". We approach this modification in two distinct ways. Using the *Detection* and *Removal* approach, we first use NLTK [2] for POS tagging to identify negation

[2] https://spacy.io/

phrases within the question. Subsequently, we systematically remove each detected negation phrase to generate modified versions of the original question. Alternatively, in the Insertion approach, we identify specific POS tags where negation can be appropriately inserted based on grammatical rules. This allows us to introduce negation phrases into questions that lack them, thereby challenging their answerability. For example, consider the question: *How much damage does breathing oxygen in space conditions cause?*. Applying the Insertion approach could yield modification such as *How much damage doesn't breathing oxygen in space conditions cause?*. Similarly, for a question like *Beyoncé does not create which aspect of her music?*, utilizing the Detection and Removal approach might lead to a question such as *Beyoncé does create which aspect of her music?*. These methods ensure that modified questions remain coherent and contextually appropriate, effectively challenging their answerability.

Mutual Exclusion. To create mutually exclusive types, we modify an answerable question such that a word or phrase becomes mutually exclusive relative to its correct answer within the context, rendering the question unanswerable. Although this method offers numerous ways to generate unanswerable questions, one strategy adopted in previous research [29] involves posing questions that demand detailed information not present within the context. By structuring questions to ask for precise details that exceed the information available in the given context, the question becomes inherently unanswerable. For instance, consider the question *When did Destiny's Child get their star on the Hollywood Walk of Fame?* which has the answer *March 2006* from the context. If we modify it to *When on March 2006 did Destiny's Child get their star on the Hollywood Walk of Fame?*, we essentially compel the context to list the specific date in March 2006, which does not exist in the context. This modification renders the question unanswerable by requesting specific details not initially provided in the context. This method ensures that the modified question remains closely aligned with the original question that is answerable, maintaining relevance to the context while challenging its ability to be answered directly.

No Information. Similar to [33], to modify the original answerable questions by considering this cause, we change the context of the question, providing no possible information for the corresponding question. In order to modify these question-context pairs, instead of modifying the original answerable questions we replaced the corresponding context with another random context from the same topic in the MRC dataset, which ensures the question is conceptually relevant to the new context. For example, with the question *Which city is the most populous in California?*, the original context provided the answer: *Los Angeles is the most populous city,..* making the question answerable. We modified the context by replacing it with an irrelevant piece of information, such as: *Southern California is also home to a large homegrown surf and skateboard culture....* This method ensures that each pair of question and modified context is grammatically correct and clearly understandable.

2.2 Unanswerable Question Evaluation

Given an answerable question $q_i \in Q_a$ as an input, this component aims to evaluate the candidate unanswerable questions C_{q_i} generated by the unanswerable question generation component, in order to select the *Final Unanswerable Question* U_{q_i} for q_i. To assess the answerability of each question within its context, we utilize six advanced MRC models described in Section 3, which are proficient in distinguishing between answerable and unanswerable questions.

Specifically, for each triple (q_i, p_i, a_i) consisting of an answerable question q_i, its associated context p_i, and the golden truth answer a_i, and for each candidate unanswerable question $(c_j, l_j) \in C_{q_i}$, we conduct the following evaluations: First, we feed each MRC model with the pair (q_i, p_i) and evaluate its extracted answer against the golden truth answer a_i using the Exact Match criterion. If the model's output matches a_i, we label it as 1; otherwise, as 0. Next, we apply the same model to (c_j, p_i): If the model identifies c_j as unanswerable, we label it as 0; otherwise, as 1. Consequently, for each pair (q_i, c_j), every MRC model generates a pair of labels from $\{(1, 1), (0, 0), (1, 0), (0, 1)\}$. We repeat this process across all six MRC models and classify (c_j, l_j) as unanswerable based on majority voting, specifically when at least 4 out of 6 models produce the label pair $(1, 0)$. We then add c_j as the unanswerable version of q_i, denoted as q'_i, to U_{q_i}, and attribute l_j as the reason for the unanswerability of q'_i.

The underlying rationale of our approach lies in the reliability of models that can accurately extract answers from the input question q_i and correctly identify candidates C_{q_i} as unanswerable within their respective contexts. This ensures a robust selection of unanswerable questions while maintaining high confidence in their validity.

3 UnAnswGen Dataset

To create the *UnAnswGen* dataset, we utilized the workflow outlined in Section 2, applying it to the SQuAD2.0 dataset as described in [29]. SQuAD2.0 is a prominent MRC dataset that pairs questions with their corresponding contexts– paragraphs from Wikipedia pages– and includes answers and the labels of whether questions are answerable. This dataset, developed through crowdsourcing, consists of a training set with 130,319 instances including 86,821 answerable and 43,498 unanswerable questions. Additionally, it features a development set comprising 11,874 instances consisting of 5,929 answerable and 5,945 unanswerable questions.

Using the Question Generation step outlined in Section 2.1, each answerable question in SQuAD2.0 undergoes a series of modifications (i.e., entity swap, number swap, antonym, negation, mutual exclusion, and no information) to generate a diverse set of Candidate unanswerable Questions (C_{q_i}). Each answerable question thus yields at least six altered versions, with an average of more than 10 variants due to the potential to generate multiple unanswerable candidate questions from a single modification process. Consequently, from the 86,821 answerable questions in SQuAD2.0, we have produced a substantial total of 944,326 candidate unanswerable questions. The distribution of these C_{q_i} across the different categories is detailed in Table 2.

In the proposed Question Evaluation module, we utilized six different existing MRC models known for their high accuracy in identifying unanswerable questions. Our selection process began with a comprehensive review of recent literature to pinpoint models that not only exhibit high performance but are also publicly available. From this review, we chose models such as Retro-Reader and

Table 2: Outcome of unanswerable question generation.

Method	# of Candidate Questions
Entity Swap	389,331
Number Swap	25,289
Negation	162,185
Antonym	324,814
Mutual Exclusion	42,707
No Information	86,820
Total	944,326

Table 3: Performance of MRC models used in unanswerable question evaluation step.

MRC Models	All questions			Answerable			Unanswerable		
	EM	ACC	F1	EM	ACC	F1	EM	ACC	F1
Retro-Reader	79.6	93.2	93.2	65.8	93	96.4	93.3	93.3	96.5
SG-Net	69.6	80	80	65.4	86.5	92.7	73.7	73.7	84.9
mdeberta-v3-base- squad2	75	85.7	85.7	68.9	90.4	95	81	81	89.5
avishkaarak-ekta-hindi	71.4	81.6	81.5	69.8	90.3	94.9	72.9	72.9	84.3
electra-base-squad2	74.8	84.7	84.7	67.9	87.8	93.5	72.2	81.6	89.9
roberta-large-squad	78.7	90	90	69.3	92.7	96.2	88	88	93.6

SG-Net for their specific capabilities. The Retro-Reader model [38], employs a two-stage verifier after extracting the answer span from the context, enhancing its ability to detect unanswerable questions. It has been fine-tuned on both SQuAD2.0 and NewsQA datasets, which include a mix of answerable and unanswerable questions, thus providing a robust testing ground for its capabilities. On the other hand, SG-Net [37], fine-tuned exclusively on the SQuAD2.0 dataset, is a neural network-based model that enhances reading comprehension through a syntax-aware self-attention mechanism. These models are designed to effectively identify unanswerable questions by returning a null or empty string when no appropriate answer is found within the context. We further select four additional models that have been specifically fine-tuned on the SQuAD2.0 dataset for MRC tasks and are publicly available via Huggingface or Github repository, namely (1) mdeberta-v3-base-squad2[3] (2) electra-base-squad2[4] (3) roberta-large-squad2 [5] and (4) avishkaarak-ekta-hindi [6]. These models are distinguished by their optimal performance characteristics, as they are designed to provide answers only when the question is definitively answerable. If a question lacks sufficient information within the given context to formulate a clear answer, these models classify the question as unanswerable. All of these four MRC models generate a confidence score indicating the model's certainty in its provided answer. For unanswerable questions, these models generate very low confidence scores, indicating the inaccuracy of any extracted answer and confirming the question's unanswerability. The performance of all six models, in terms of F1 scores and accuracy, is summarized in Table 3.

Finally, we employ a majority voting mechanism on C_{q_i} to refine and determine the final unanswerable question set. The approach that is used to generate each unanswerable question (described in section 2.1) is recorded as its label. Table 4 provides a detailed

[3]https://huggingface.co/timpal0l/mdeberta-v3-base-squad2
[4]https://huggingface.co/deepset/electra-base-squad2
[5]https://huggingface.co/deepset/roberta-large-squad2
[6]https://huggingface.co/AVISHKAARAM/avishkaarak-ekta-hindi

Table 4: Statistics on UnAnswGen dataset.

Unanswerability Classes	# of Questions	Percentage	Average proportion
Entity Swap	17,444	14.77	20.09
Number Swap	2,255	1.90	2.6
Negation	45,053	38.06	51.89
Antonym	27,749	23.44	31.95
Mutual Exclusion	3,221	2.72	3.71
No Information	22,652	19.14	26.08
Total	118,374		

breakdown of the *UnAnswGen* dataset's composition by showcasing the distribution of unanswerable questions across different categories of unanswerability, which totals 118,374 questions. It also includes the percentage of questions that were augmented in each category. According to the data presented in Table 4, those C_{q_i} that derived from three specific unanswerability categories (i.e., Negation, Antonym, and No Information) constitute the majority of the final unanswerable question set. Specifically, questions from the Negation category account for 38.06% of the final set, those from Antonym account for 23.44%, and those from No Information represent 19.04%. Conversely, the categories of Number Swap and Mutual Exclusion contribute the least to the final unanswerable question set. Another observation from the *UnAnswGen* analysis is the average number of unanswerable questions generated in each category relative to each answerable question (indicated as average proportion in Table 4). We observe that within the *UnAnswGen* dataset, each answerable question is significantly augmented with unanswerable counterparts through various modifications. Specifically, an average of 51.89 unanswerable questions are generated under the Negation label, 31.95 unanswerable questions under the Antonym label, and only 2.6 unanswerable questions per answerable question with the Number Swap label.

4 Human Assessments

We conducted an additional human evaluation to assess the effectiveness of our methods for generating unanswerable questions, using the following three criteria adapted from [39]:

Unanswerability. This criterion measures whether the generated question can be answered based on the provided context. A score of 0 indicates that the question is easily answerable, while a score of 1 signifies that the question is designed to be unanswerable.

Contextual Relevance. This criterion assesses how closely the generated question relates to the context provided. A score of 0 indicates that the question is completely unrelated to the context, whereas a score of 1 indicates strong relevance to the context.

Clarity and understanding. This criterion evaluates the clarity and coherence of the generated question. A score of 1 indicates that the question is incomprehensible, 2 suggests minor errors that do not significantly affect the meaning, and 3 reflects a clear and coherent question structure. For this experiment, we randomly selected 20 unanswerable questions from each class: *Entity Swap*, *Number Swap*, *Negation*, *Antonym*, *Mutual Exclusion*, and *No Information*. These questions were drawn from the final unanswerable question set, including their corresponding contexts, original answerable questions, and respective labels. In total, our evaluation comprised 120 questions. Three experts assessed the random sample

Table 5: Performance of three models fine-tuned on SQuAD2-CR and SQuAD2-CR+UnAnswGen.

Model	Dataset	All		Answerable		Unanswerable		Entity Swap		Number Swap		Antonym		Negation		Mutual Exclusion		No Info	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
RoBERTa	SQuAD2-CR	71.83	71.83	91.41	95.51	52.31	59.24	48.91	65.69	72	83.73	62.41	76.86	86.43	92.72	**12.22**	**21.78**	13.59	23.92
	SQuAD2-CR+UnAnswGen	**73.7**	**73.7**	**92.41**	**96.05**	**55.04**	**60.95**	**50.83**	**67.4**	**73.83**	**84.95**	**69.34**	**81.89**	**89**	**94.18**	10.47	18.96	**15.27**	**26.49**
DeBERTa	SQuAD2-CR	70.54	70.54	**93.81**	**96.81**	47.32	55.57	39.77	56.9	68.56	81.35	58.7	73.97	86.43	92.72	12.97	22.96	**10.53**	**19.06**
	SQuAD2-CR+UnAnswGen	**71.93**	**71.93**	92.48	96.09	**51.43**	**58.27**	**46.83**	**63.78**	**69.17**	**81.77**	**64.86**	**78.69**	**86.8**	**92.93**	**13.47**	**23.74**	9.77	17.8
Electra	SQuAD2-CR	73.65	73.65	**93.57**	**96.68**	53.78	60.15	**51.96**	**68.39**	70.79	82.9	61.74	76.34	88.63	93.97	**15.21**	**26.41**	13.28	23.45
	SQuAD2-CR+UnAnswGen	**74.02**	**74.02**	93.31	96.54	**54.79**	**60.39**	51.55	68.03	**76.27**	**86.54**	**64.86**	**78.69**	**89**	**94.18**	14.96	26.03	**13.89**	**24.4**

Table 6: Inter annotator agreement.

	Unanswerability	Relatedness	Readability
Krippendorff's α	0.86	0.86	0.85

Table 7: Different training sets used in our benchmark.

Category	SQuAD2-CR	SQuAD2-CR+UnAnswGen	Δ
Entity Swap	17,143	17,143	0
Number Swap	4,570	6,825	2,255
Negation	5,647	17,647	12,000
Antonym	9,673	17,673	8,000
Mutual Exclusion	3,115	6,336	3,221
No Information	3,350	17,350	14,000
Answerable	86,821	86,821	0

to evaluate the UnAnswerGen dataset against these criteria. Table 6 presents the results of Krippendorff's α metric [9], which measures agreement among all annotators' assessments. Importantly, all experts agreed unanimously on the questions' unanswerability, relevance to the context, and readability.

5 Establishing Benchmarks on UnAnswGen Dataset

To establish a benchmark and assess the effectiveness of our UnAnswGen dataset, we evaluated its impact on the downstream task of predicting the cause of unanswerability. Our UnAnswGen dataset, detailed in Section 3, consists of unanswerable questions across six categories, all derived from the answerable questions in the SQuAD2.0 dataset. We conducted benchmark tests by fine-tuning a model on two versions of the training set: (1) the original SQuAD2-CR, which already includes multi-class labeling of unanswerable questions, and (2) the SQuAD2-CR training set enriched with the UnAnswGen dataset.

Due to a significant class imbalance in the original SQuAD2-CR dataset (see Table 7), we augmented it with selections from our UnAnswGen dataset to create a more balanced version, termed SQuAD2-CR+UnAnswGen. This integration addresses the low-data regime issue, ensuring a more uniform distribution across all unanswerability categories, as detailed in Table 7.

For the multi-classification task of assigning an unanswerability type to input question-context pairs, we utilized several advanced pre-trained large models: based version of RoBERTa [23], small version of DeBERTa [10], and base version of Electra [5]. All models underwent training on both the enhanced SQuAD2.0+ UnAnswGen and the original SQuAD2-CR datasets, with each model fine-tuned for 3 epochs using a learning rate of 2e-5, a batch size of 4, and a maximum sequence length of 512.

Table 5 presents a performance comparison of three advanced transformer models—RoBERTa, DeBERTa, and Electra—fine-tuned on two versions of the datasets. Notably, while the answerable questions in SQuAD2-CR remain unchanged in both datasets, models trained on the balanced dataset demonstrate slight improvements in accuracy and F1 scores (1-2%) across all categories when compared to those trained on the imbalanced dataset. These performance enhancements are particularly evident in unanswerable question categories where accuracy gains of 3% for RoBERTa, 4% for DeBERTa, and 1% for Electra were observed. The balanced dataset successfully mitigates issues related to the low-data regime, resulting in enhanced model performance for specific unanswerable categories such as Number Swap, Antonym, and Negation, and shows improvements in categories like Entity Swap, Mutual Exclusion, and No Information across most models.

Overall, the balanced augmentation of the SQuAD2-CR dataset not only boosts overall model performance but also enhances their capability to handle various types of unanswerable questions, underlining the importance of balanced data in training robust transformer models that aim to predict the causes of unanswerability.

6 Conclusion

In this paper, we propose a configurable workflow to generate enhanced MRC datasets, with a focus on including multi-label unanswerable questions. We have developed the UnAnswGen dataset, which features a broad spectrum of unanswerable questions. This dataset, along with the source code for our workflow, has been made publicly available to support the MRC research community. Additionally, we conducted benchmark tests to demonstrate the effectiveness of our UnAnswGen dataset in predicting the cause of unanswerability using state-of-the-art models in a multi-class classification task. The results from these benchmark tests reveal that our dataset significantly enhances the models' ability to accurately classify different types of unanswerability.

In the future, we intend to extend the application of our workflow to enrich other datasets, such as HotPotQA [35] and Natural Questions [16], with multi-label unanswerable questions. Another promising avenue is to further develop capabilities that not only enhance the detection of unanswerable questions but also empower researchers to create solutions, such as query reformulation. This would involve transforming unanswerable questions into answerable ones based on their causes, thereby significantly improving the interaction between users and question-answering systems.

References

[1] Seohyun Back, Sai Chetan Chinthakindi, Akhil Kedia, Haejun Lee, and Jaegul Choo. 2020. NeurQuRI: Neural question requirement inspector for answerability prediction in machine reading comprehension. In *International Conference on Learning Representations*.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[3] Chang Nian Chuy, Qinmin Vivian Hu, and Chen Ding. 2023. One Stop Shop for Question-Answering Dataset Selection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 3115–3119.

[4] Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723* (2017).

[5] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

[6] Christiane Fellbaum. 1998. *WordNet: An electronic lexical database.*

[7] Ge Gao, Hung-Ting Chen, Yoav Artzi, and Eunsol Choi. 2023. Continually Improving Extractive QA via Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 406–423.

[8] Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. 2023. A Lightweight Method to Generate Unanswerable Questions in English. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7349–7360.

[9] Kilem L Gwet. 2011. On the Krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October* 2, 2011 (2011), 2011.

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).

[11] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*. 37–46.

[12] Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6529–6537.

[13] Kevin Huang, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Relation module for non-answerable predictions on reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 747–756.

[14] Yunjie Ji, Liangyu Chen, Chenxiao Dou, Baochang Ma, and Xiangang Li. 2022. To answer or not to answer? Improving machine reading comprehension model with span-based contrastive learning. *arXiv preprint arXiv:2208.01299* (2022).

[15] Souvik Kundu and Hwee Tou Ng. 2018. A nil-aware answer extraction framework for question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4243–4252.

[16] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[17] Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. SQuAD2-CR: Semi-supervised annotation for cause and rationales for unanswerability in SQuAD 2.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 5425–5432.

[18] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115* (2017).

[19] Jinzhi Liao, Xiang Zhao, Jianming Zheng, Xinyi Li, Fei Cai, and Jiuyang Tang. 2022. Ptau: Prompt tuning for attributing unanswerable questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1219–1229.

[20] Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Jiancheng Lv, Nan Duan, and Ming Zhou. 2020. Tell Me How to Ask Again: Question Data Augmentation with Controllable Rewriting in Continuous Space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5798–5810.

[21] Qian Liu, Rui Mao, Xiubo Geng, and Erik Cambria. 2023. Semantic matching in machine reading comprehension: An empirical study. *Information Processing & Management* 60, 2 (2023), 103145.

[22] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic Answer Networks for Machine Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1694–1704.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[24] Dheeraj Mekala, Jason Wolfe, and Subhro Roy. 2023. ZEROTOP: Zero-Shot Task-Oriented Semantic Parsing using Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5792–5799.

[25] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).

[26] Hariom A Pandya and Brijesh S Bhatt. 2021. Question answering survey: Directions, challenges, datasets, evaluation matrices. *arXiv preprint arXiv:2112.03572* (2021).

[27] Wei Peng, Yue Hu, Jing Yu, Luxi Xing, and Yuqiang Xie. 2021. APER: adaptive evidence-driven reasoning network for machine reading comprehension with unanswerable questions. *Knowledge-Based Systems* 229 (2021), 107364.

[28] Rifki Afina Putri and Alice Oh. 2022. IDK-MRC: Unanswerable Questions for Indonesian Machine Reading Comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 6918–6933.

[29] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

[30] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[31] Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1683–1693.

[32] Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Weifeng Lv, and Ming Zhou. 2018. I know there is no answer: Modeling answer validation for machine reading comprehension. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7*. Springer, 85–97.

[33] Son Quoc Tran, Gia-Huy Do, Phong Nguyen-Thuan Do, Matt Kretchmar, and Xinya Du. 2023. AGent: A Novel Pipeline for Automatically Creating Unanswerable Questions. *arXiv preprint arXiv:2309.05103* (2023).

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

[35] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2369–2380.

[36] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences* 10, 21 (2020), 7640.

[37] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. SG-Net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9636–9643.

[38] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14506–14514.

[39] Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to Ask Unanswerable Questions for Machine Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4238–4248.