Reference-based Metrics Disprove Themselves in Question Generation

Anonymous ACL submission

Abstract

Reference-based metrics such as BLEU and BERTScore are widely used to evaluate question generation (QG). In this study, on QG benchmarks such as SQuAD and HotpotQA, we find that using human-written references cannot guarantee the effectiveness of the reference-based metrics. Most QG benchmarks have only one reference; we replicated the annotation process and collect another reference. A good metric was expected to grade a human-validated question no worse than generated questions. However, the results of reference-based metrics on our newly collected reference disproved the metrics themselves. We propose a reference-free metric consisted of multi-dimensional criteria such as naturalness, answerability, and complexity, utilizing large language models. These criteria are not constrained to the syntactic or semantic of a single reference question, and the metric does not require a diverse set of references. Experiments reveal that our metric accurately distinguishes between high-quality questions and flawed ones, and achieves state-of-the-art alignment with human judgment.

1 Introduction

Question generation (QG) usually refers to the task of answer-aware question generation for controllability, aiming at generating a question based on a given context and answer span. Solutions are used to improve educational tools, build a product-based question-answering (QA) database, etc. Though anchored on a specific answer, there are still multiple ways of framing a question semantically and syntactically (Yu and Jiang, 2021; Cho et al., 2019). Users expect quality of every generated question.

To evaluate QG performance, reference-based metrics are widely used, which assess a machine-generated question against a human-written reference. The metrics are calculated either at the word level such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), or in the embedding space such as BERTScore (Zhang et al., 2019). The challenges of using these evaluation metrics speak to the metrics themselves, considering word overlaps and/or semantic similarity between the generated question and the reference. In this sense, a QG model can "cheat" on the metrics by using many similar words to the reference, but ignoring essential components of a question. Mohammadshahi et al. questioned the effectiveness of the reference-based metrics, developed a QA model, and defined a new metric named "answerability" or RQUGE. Though they showed a slightly higher correlation with human preference, the failure of reference-based metrics was not studied, and the new metric's effectiveness is sensitive to the QA model's training and limited to its ability.

To disprove existing metrics, the challenge can be traced to the lack of diverse references for benchmark datasets. Previous works have shown that with access to a more diverse pool of references, the problem of poor correlation for these metrics can be mitigated (Freitag et al., 2020; Tang et al., 2023). However, QG benchmarks often contain one human-written ground-truth per example.

Our study starts from collecting another set of human-written references for two QG benchmarks, following their standard annotation instructions. We also collect three groups of candidate questions, each lacking in an essential aspect of a question, for comparison. We study how five reference-based metrics, namely BLEU-4 (Papineni et al., 2002), BLEURT (Sellam et al., 2020), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), and Q-BLEU (Nema and Khapra, 2018), and two reference-free metrics, QAScore (Ji et al., 2022), and RQUGE (Mohammadshahi et al., 2023), score the four groups of questions. Figure 1 highlights the incompetency of current QG metrics in distinguishing the new reference (a valid question;

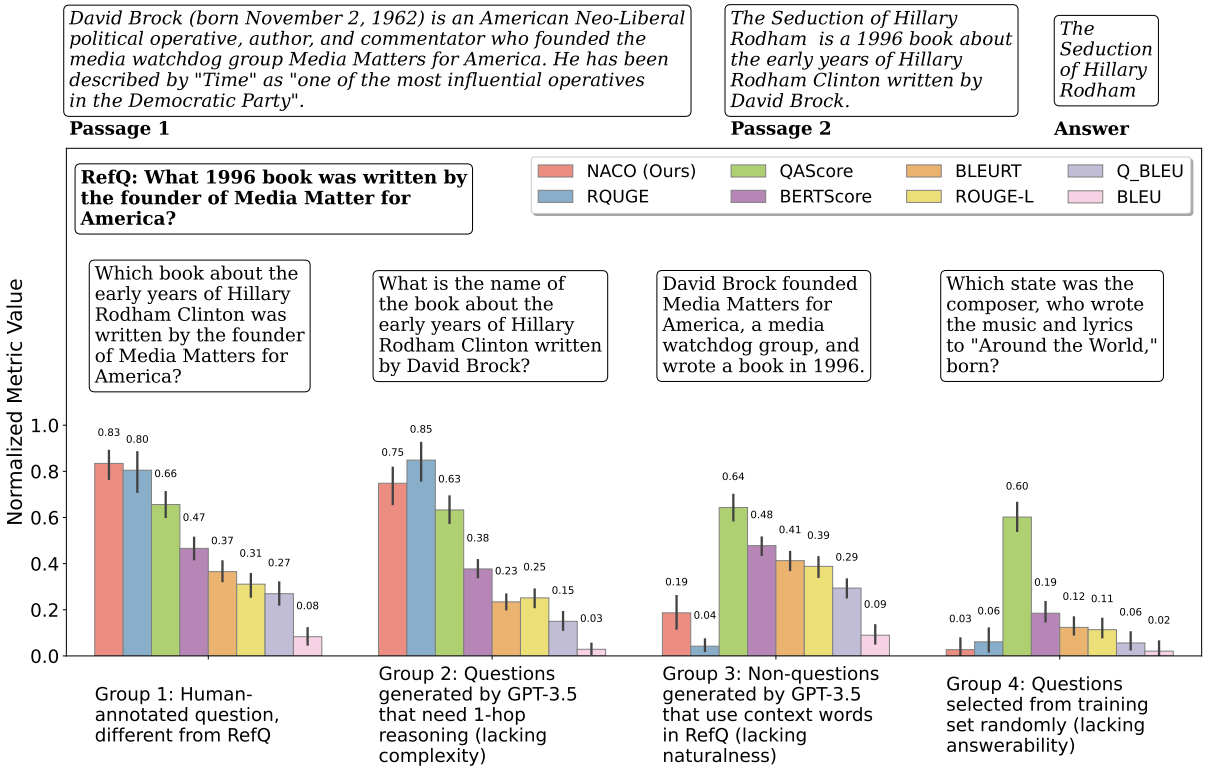

Figure 1: Normalized value of different evaluation metrics for four types of candidate questions against the same reference (RefQ) in the HotpotQA dataset (Yang et al., 2018). A good metric should give a significantly higher score to Group 1 than the other three groups. It can be seen that current reference-based metrics (in colors except red and blue) can identify the random questions (Group 4). However, they have trouble separating questions from Group 1 and Group 3. RQUGE, although capable of identifying two groups of questions that violate naturalness (Group 3) and answerability (Group 4), assigns a higher score for Group 2, which lacks complexity, than for Group 1. Our proposed metric (in red), prioritizing essential criteria of a question, can effectively distinguish all four groups of candidates while maintaining the highest rating for the valid questions.

see Group 1) from a less-complex-than-referenced question (Group 2), a non-question sentence that uses similar words (Group 3) or a randomly-selected question from training set (Group 4). Although these metrics tend to give higher scores for the new references than random questions, it remains challenging to separate them from the other less desirable candidates.

Based on the above observations, we assert the failure of reference-based metrics in QG evaluation. We propose a shift to an evaluation mechanism that addresses essential criteria of a question that current metrics do not: (1) **N**aturalness: *how natural the question sound* (Wang et al., 2020; Bi et al., 2021), (2) **A**nswerability: *whether the question is grounded to the given answer* (Ushio et al., 2022; Ji et al., 2022; Nema and Khapra, 2018; Mohammadshahi et al., 2023), and (3) **Co**mplexity: *how likely it requires inferencing and synthesizing information* (Wang et al., 2020; Bi et al., 2021). These criteria are not constrained to the syntactic and semantic structure of a single reference question. Thus, they can help tackle the challenges of evaluating question quality without access to a diverse set of references.

To overcome the limitation of the answerability measure in RQUGE and implement the other two measures, we utilize large language models (LLMs), which have demonstrated potential utility in data annotation tasks (Liu et al., 2023; Wang et al., 2023; Lin and Chen, 2023; Chiang and Lee, 2023), and their Chain-of-Thought (CoT) (Wei et al., 2022) process. We design CoT prompts for the LLM to directly measure the three criteria, as described in detail in §3.

We name the three-dimensional metric **NACo**. The red bars (the leftmost in each group) in Figure 1 show that NACo successfully distinguishes the valid questions (i.e., new human-written reference) from the other three groups with significant margins. Reference-based metrics are so heavily influenced by the presence of overlapping words

between the original reference and an invalid candidate that they even prefer the invalid candidate that NACo assigns a significantly lower score.

The key contributions of this paper include:

- We produce an additional set of human-written questions to current QG benchmarks, and show the unreliability of reference-based metrics in reflecting the quality of generated questions.
- We propose NACo, a novel evaluation metric bridging the gap between human assessment and automated evaluation by assigning scores to three criteria of a good question.
- Through experiments and human evaluation, we demonstrate that NACo better aligns with human judgment of a good question than reference-based metrics for QG.

2 Failure of Reference-based QG Metrics

2.1 Data Collection

In this study, we question the reliability of current evaluation metrics for question generation. To do so, we collect a new set of valid human-written questions and three other sets of questions for benchmark datasets, each violating an important criterion of a question addressed by NACo. Figure 1 illustrates our data collection process.

Group 1: Human-written questions qualified as another reference for benchmark datasets: We follow the procedure adopted by most papers collecting QA datasets. For each example to be annotated, we ask annotators, all fluent English speakers, to create a question based on some context passage(s) and a given answer (Rajpurkar et al., 2016). If two passages are provided, we ask annotators to create a question such that it requires reasoning over both passages (Yang et al., 2018).

Liu et al. proposed a concept of *clues* for QG, which refers to words from the context passage that also appear in the question. Their experimental results indicate that the addition of a clue-prediction model enhances the performance of question generators on reference-based metrics. We investigate the usefulness of this concept by asking the annotators to phrase an additional question such that it contains the clue words used by the original annotators of the datasets. We ensure that the clue words are only presented to the annotators after they have finished asking their first question.

We perform the additional annotation on two datasets: (1) 748 test examples of SQuAD (Rajpurkar et al., 2016), and (2) 48 test examples of HotpotQA (Yang et al., 2018). For the HotpotQA sample, we also collected three other sets of questions, each violating an aspect required by the reference questions.

Group 2: Single-hop questions for a multi-hop QG benchmark: This group of candidate questions targets the multi-hop characteristic of HotpotQA where the ground-truth questions are formed based on two passages. Specifically, we selected one from the original two passages that contains the answer span. We then asked GPT-3.5 to generate a question based on this single context passage. Next, we reviewed the questions for grammar, clarity, relevance to the passage, independence from external knowledge, and a logical path to the answer. We hypothesize that current QA-based metrics such as RQUGE (Mohammadshahi et al., 2023) and QAScore (Ji et al., 2022) fail to prefer the two-hop human-written questions to this set of single-hop questions.

Group 3: Non-questions that use the same words as the reference: For this group of questions, we asked GPT-3.5 to generate a sentence based on the passages and use as many words from the same list of clues given to our annotators. We added a constraint such that the generated sentence cannot be in the form of a question. We then manually went through the generated sentences to ensure that no hallucinations were in place. In this sense, we produced a group of candidates that does not satisfy the most basic linguistic requirement of a question, naturalness, but still manages to contain many similar words as the ground-truth questions. We hypothesize that the majority of current evaluation metrics, especially those that rely mainly on word overlaps, fail to distinguish between the valid human-written questions in Group 1 and this group of non-questions.

Group 4: Random questions from the training set: The final set of candidate questions comes randomly from the training set of the benchmark. In the example illustrated in Figure 1, the random question cannot be grounded to the given answer *The Seduction of Hillary Rodham*. Specifically, the answer to this candidate question is *Robin McLaurin Williams*, which is completely irrelevant to the given answer. In this sense, this group of candidate questions violates the answerability aspect of an ideal candidate. We hypothesize that the majority of current evaluation metrics can effectively recognize this set of random questions.

2.2 Study Design and Results

Previous studies aiming to challenge the effectiveness of reference-based metrics in QG typically rely on human evaluation. That is, they investigate whether the scores given to generated questions by QG metrics are highly correlated with the scores given by human evaluators (Mohammadshahi et al., 2023; Ji et al., 2022). Unlike these studies, our research adopts a different approach during the data collection phase for benchmark datasets. Specifically, we replicate the data collection procedure of SQuAD and HotpotQA to collect new references, referred to as Group 1. Our focus is on determining if the newly collected references, when evaluated as candidates against the original references, receive high ratings from existing reference-based metrics. For the HotpotQA dataset, we extended our collection to include three additional groups of candidate questions considered less desirable (groups 2, 3, and 4) to ensure comprehensive comparisons. An effective and robust evaluation metric should assign a significantly higher score for questions in Group 1 compared to those in other groups.

Figure 1 shows the average normalized scores given by reference-based metrics to the four groups of candidate questions, all based on the same references. We find that all reference-based metrics, BLEU, ROUGE-L, BLEURT, Q-BLEU, and BERTScore, can effectively distinguish Group 4 (random questions) from the other groups, assigning it significantly lower score. For instance, the average ROUGE-L score for Group 4 is 0.11, compared to 0.31 for Group 1, 0.25 for Group 2, and 0.39 for Group 3, with a minimum difference of 14% from the scores of the other groups.

Figure 1 also reveals issues with the reference-based metrics in accurately assessing Groups 1, 2, and 3. Notably, for all five reference-based metrics, Group 3, non-question sentences with wording similar to the references, receives the highest average score. For example, the ROUGE-L metric scores a non-question sentence that uses similar wording to the reference on average 8% higher than a new reference produced by our annotators, and 14% higher than a perfectly answerable question requiring less reasoning than the reference. This indicates a flaw in reference-based metrics, as candidates that do not form coherent questions should not receive higher scores than those that do.

The recently-introduced reference-free metrics, QAScore and RQUGE, also face difficulties in giving reasonable scores to questions from Groups 1, 2 and 3. QAScore, despite rating the new references highest among four groups, shows minimal score differences. Meanwhile, RQUGE gives the highest average score (0.85) to Group 2, which contains single-hop questions in contexts requiring multi-hop reasoning. RQUGE's preference for single-hop questions can be attributed to its disregard for the complexity of the candidate question. It utilizes a pretrained QA model to compute a score based on the model's responses to the candidate question. The questions we collected, which require reasoning over two documents, may pose a greater challenge for the QA model compared to the simpler questions from Group 2. Since RQUGE's scoring mechanism does not consider the question's complexity, it underestimates the new references we collected in Group 1, scoring them at 0.80.

Given the limitations of existing reference-based and reference-free metrics in accurately evaluating the four groups of questions, we propose a novel reference-free metric. This new metric aims to assess the quality of a question across multiple dimensions, providing a broader and more nuanced framework for assessing generated questions.

3 NACo: A Novel Multi-dimensional Reference-free QG Metric

We propose NACo, which leverages prompting and Chain-of-Thought (CoT) reasoning (Wei et al., 2022) in answering the questions to be evaluated to obtain a score for each of the following criteria: **N**aturalness, **A**nswerability, and **C**omplexity. More specifically, given the relevant context passage(s) and a question, we instruct an LLM to answer the question based on the passage(s) as follows:

- The LLM first reads over the context passage(s) and the question. The LLM check whether the question makes these mistakes: (1) not a question, (2) grammar errors, or (3) unclear objective. If so, the LLM should respond with '*Question unnatural*', and we assign a score of 0 for the question in terms of **naturalness** n_{cand}. Otherwise, n_{cand} is 1.
- Next, the LLM performs CoT reasoning to answer the question. Based on the LLM's CoT response, we obtain the **complexity** of the question by counting the number of reasoning steps the LLM made to answer the question.
- The LLM provide the final answer to the question. We define the **answerability** of the ques-

tion a_{cand} as the F1 score between the LLM's answer to the question and the ground-truth answer used to generate the question.

The inherent qualities of questions speak to naturalness (Mohammadshahi et al., 2023) and answerability (Nema and Khapra, 2018; Ji et al., 2022; Mohammadshahi et al., 2023), where higher values in these criteria indicate better quality in a question. We adopt a hierarchical scoring scheme that first examines the naturalness and answerability score obtained following the CoT-QA process. If the candidate question scores 0 in these aspects, it is assigned a NACo score of 0.

If a candidate question passes the initial naturalness and answerability evaluation, we determine whether its complexity aligns with expected standards. Importantly, complexity expectations vary among domains and datasets. For instance, in the HotpotQA dataset, questions that require multi-hop reasoning might be preferred over simpler, single-hop questions. This observation may not apply for other datasets. In this sense, NACo relies on a subset of examples from the specific dataset to find the *expected complexity* of a question in that dataset. Specifically, we perform the above CoT-QA process to obtain the complexity of the references. Expected complexity is then defined by the most common number of reasoning steps needed by the LLM to answer a reference question. Our experiments using 750 examples from the training set find that for GPT-3.5, the expected complexity is 2 for questions in SQuAD, and 3 for questions in HotpotQA. Subsequently, NACo measures the similarity, denoted by c_{cand}, between the complexity of the candidate question and the expected complexity. Overall, NACo is a weighted combination of n_{cand}, a_{cand}, and c_{cand}. In our experiments, we adopt a fair weight $\frac{1}{3}$ for each criterion.

4 Experiments

4.1 Experimental Setup

We compare the evaluation capacity of NACo with that of current QG metrics on four QG models and three sets of human-validated references.

Generative Language Models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) are current state-of-the-art QG performers on reference-based metrics (Ushio et al., 2022). We fine-tune BART-base using the training set, following the method introduced by Chan and Fan. We refer to this model as **BART-base**. We also produce another version of BART-base that explores the effectiveness of clue-guided QG (Liu et al., 2020). Specifically, before finetuning, we highlight the ground-truth clues used by reference questions (RefQ) in the context given as input to BART-base. We refer to this version as **BART-clue-RefQ**. In addition, we use GPT-3.5 to generate questions for the test examples through zero-shot and few-shot prompting. In the few-shot setting, we randomly select 10 examples from the training set of the dataset as demonstrations. We refer to these models as **GPT-3.5 (zero-shot)** and **GPT-3.5 (few-shot)**.

Human-validated Questions: Alongside the original reference questions provided by the datasets (**RefQ**), we use the annotated data detailed in §2 to obtain two human-validated competitors: **AnnoQ**, which contains the questions written by our annotators before given gold clues, and **AnnoQ-clue-RefQ**, which contains the gold-clue-guided questions written by our annotators.

4.2 Results

4.2.1 Failure of reference-based metrics

We report the performance of QG competitors on different evaluation metrics, including NACo. For reference-based metrics, we report their values using RefQ as the reference in Table 1 and 2. Values of reference-based metrics when AnnoQ are used as references are available in Appendix A.3.

Even though RefQ, AnnoQ, and AnnoQ-clue-RefQ are all qualified as valid questions, a large gap exists between them when reference-based metrics are in use. Specifically, in the SQuAD dataset, using RefQ as a reference, the BLEU scores for RefQ, AnnoQ, and AnnoQ-clue-RefQ are 100, 12.78, and 27.43, respectively, as shown in Table 1. Conversely, the gap in the reference-free metrics like RQUGE and NACo among these different sets of human-validated questions is much narrower when given the same input. Table 1 presents the NACo scores for RefQ, AnnoQ, and AnnoQ-clue-RefQ in the SQuAD dataset as 75.09, 74.01, and 74.21, respectively. Similar patterns are observed in the HotpotQA dataset, as detailed in Table 2.

According to reference-based metrics, models that learn from training data either through fine-tuning (like BART-base) or demonstration (like GPT-3.5) are scored significantly higher than our annotators, who lack access to the training data. For instance, in the case of SQuAD, BART-base is scored higher than AnnoQ by almost 7% according

QG Competitor	Ref-based metrics					Ref-free metrics		Our metric			
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	Na.	Ans.	Cmp.	NACo
LM-generated											
BART-base	19.53	-0.28	44.79	92.13	36.94	**-0.37**	4.62	99.60	66.75	69.15	73.30
GPT-3.5 (few-shot)	18.06	-0.23	43.58	92.18	36.48	**-0.37**	4.56	**100.00**	65.23	69.63	73.67
BART-clue-RefQ	**31.91**	**0.07**	**59.92**	**94.37**	**52.33**	-0.38	4.56	99.73	63.94	69.24	69.97
Human-validated											
RefQ	100.00	100.00	100.00	100.00	100.00	-0.38	**4.89**	99.87	**68.67**	68.93	**75.09**
AnnoQ	12.78	-0.31	37.83	91.52	31.32	**-0.37**	4.71	99.73	67.70	70.33	74.01
AnnoQ-clue-RefQ	27.43	0.04	53.62	93.85	46.89	-0.38	4.76	99.87	66.34	**70.36**	74.21

Table 1: **SQuAD** - Performance of different QG methods on NACo and other existing metrics. The evaluation uses original SQuAD questions (RefQ) as references, with GPT-3.5 as the underlying QA system. Baseline metrics include BLEU (B), BLEURT (B-RT), ROUGE-L (R-L), BERTScore (BSc), Q-BLEU (Q-B), QAScore (QA-S), and RQUGE (R-Q). We also report individual score for each criterion Naturalness (Na.), Answerability (Ans), and Complexity (Cmp.) as measured by our NACo metric. The notation introduced here is consistent across subsequent tables for clarity and ease of comparison. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.

QG Competitor	Ref-based metrics					Ref-free metrics		Our metric			
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	Na.	Ans.	Cmp.	NACo
LM-generated											
BART-base	12.97	-0.79	34.38	87.88	27.67	-0.28	2.78	64.58	31.00	68.56	41.25
GPT-3.5 (zero-shot)	7.94	-0.95	26.05	87.38	17.00	-0.28	4.25	**100.00**	71.55	70.72	73.60
GPT-3.5 (few-shot)	10.53	-0.87	27.51	87.76	18.67	**-0.27**	4.21	99.31	67.31	72.64	71.09
BART-clue-RefQ	31.90	-0.10	**58.77**	**92.61**	**51.98**	-0.28	3.31	90.28	46.84	79.41	63.17
Human-validated											
RefQ	100.00	100.00	100.00	100.00	100.00	**-0.27**	4.21	99.31	59.62	**83.02**	74.33
AnnoQ	13.64	-0.70	30.94	89.22	26.95	**-0.27**	4.32	**100.00**	**76.11**	82.05	**83.43**
AnnoQ-clue-RefQ	**35.59**	**-0.06**	52.95	92.15	51.76	**-0.27**	**4.44**	**100.00**	71.46	81.46	81.60

Table 2: **HotpotQA** - Performance of different QG methods on NACo and other existing metrics. The evaluation uses original HotpotQA questions (RefQ) as references, with GPT-3.5 as the underlying QA system. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.

to BLEU-4, reported in Table 1. As reference-based metrics measure syntactic and semantic similarity, the use of a single reference can disqualify our annotated questions from being considered reference materials, resulting in a misleading portrayal of a valid group of candidate questions.

4.2.2 Effectiveness of NACo

In this section, we demonstrate NACo's superiority over both current reference-based and reference-free metrics. Referring to our analysis of four groups of candidate questions for HotpotQA, as illustrated in Figure 1, NACo uniquely succeeds in separating all four groups by significant margins, unlike the seven existing metrics. The newly collected multi-hop questions in Group 1, which meet all critical criteria for HotpotQA questions, achieve the highest average NACo score of 0.83. They are followed by the questions in Group 2, lacking in complexity, with a score of 0.75; Group 3, lacking in naturalness, with a score of 0.19; and Group 4, lacking in answerability, with a score of 0.03.

Table 2 shows that reference-based metrics rate BART-base questions slightly lower than AnnoQ (by 0.67%). On the other hand, this gap is significantly larger when NACo is in use (42.18%). Upon manually reviewing the questions generated by BART-base, we noticed a considerable number of them were not actual questions but rather statements using similar wording to the reference question RefQ. Figure 2 presents a case study illustrating this issue. Here, BERTScore, the reference-based metric with highest alignment with human judgment (Mohammadshahi et al., 2023), prefers the text generated by BART-base—even though it

Context and Answer:
Passage 1: " Lari Michele White (; born May 13, 1965) is an American country music artist and actress. She first gained national attention in 1988 as a winner on "You Can Be a Star", […]
Passage 2: "I Will Not Say Goodbye" is a song written by Lari White, Chuck Cannon and Vicky McGehee, and recorded by "American Idol" season 8 finalist Danny Gokey. […]

- **RefQ:** "I Will Not Say Goodbye" is a song written in part by a music artist who first gained national attention as a winner of what talent competition?
- **AnnoQ**: Which 1988 competition did a co-writer of "I Will Not Say Goodbye" become a winner of? *NACo: 88.89; BERTScore: 49.37*
- **BART-base**: "I Will Not Say Goodbye" is a song written by Chuck Cannon and Vicky. *NACo: 0; BERTScore: 54.48*

Figure 2: Case study 1: NACo vs BERTScore. Longest common subsequences between candidate question and RefQ are highlighted.

Context and Answer:
Passage 1: "The Guadalcanal Campaign, also known as the Battle of Guadalcanal and codenamed *Operation Watchtower* was a military campaign fought between 7 August 1942 and 9 February 1943 on and around the island of Guadalcanal in the Pacific theater of World War II […]
Passage 2: Joseph Jacob "Joe" Foss (April 17, 1915 – January 1, 2003) was a United States Marine Corps major […] He received the Medal of Honor in recognition of his role in air combat during the Guadalcanal Campaign.

- **RefQ:** What was the codename of the campaign where Joe Foss received a Medal of Honor? *NACo: 87.96; RQGUE: 93.17*
- **GPT3.5 (zero-shot):** What was the codename for the military campaign fought between 7 August 1942 and 9 February 1943 on and around the island of Guadalcanal in World War II? *NACo: 81.48; RQGUE: 94.52*

Figure 3: Case study 2: NACo vs RQUGE. Context words used by the question are highlighted in the same color if they come from the same passage.

does not form a question—over the newly collected reference. This incompetency of reference-based metric can be explained by the fact that BART, when finetuned on the HotpotQA training set, can identify words that will be used in the reference RefQ, without necessarily forming a coherent and answerable question. NACo, emphasizing essential criteria of a question, assigns a score of 0 to the BART-base output while giving a high score for AnnoQ (88.89). This finding aligns with our study findings in §2, which contrasted Group 1 (new references) with Group 3 (non-questions using the same words as the reference).

Table 2 also shows that the recent reference-free metric for QG - RQGUE - rates questions generated by GPT-3.5, in both zero-shot or few-shot settings, approximately the same as the original reference question (RefQ). Our manual investigation revealed that GPT-3.5 often makes use of only one of the two given context passages, despite being instructed to generate a multi-hop question, as demonstrated in a case study in Figure 3. Specifically, all context words used by GPT-3.5 in the case study are drawn exclusively from Passage 1, making access to Passage 2 unnecessary for answering the question. Meanwhile, RefQ incorporates context words from both passages, crafting a question that requires reasoning across both for an answer. RQUGE assigns a higher score for the GPT-3.5 question than for RefQ (94.52 and 93.17, respectively). This parallels our findings in §2, which evaluated how current QG metrics assess Group 1 (new references we collected) and Group 2 (single-hop questions in a multi-hop context), highlighting a failure to consider question complexity. RQUGE, utilizing a pretrained QA model for evaluation, overlooks this aspect. Addressing this gap, NACo acknowledges the answerability and naturalness of the GPT-3.5 question, but penalizes it for its lower-than-expected complexity, resulting in a score of 81.48. Since RefQ meets all three criteria of a candidate question, NACo awards it a higher score of 87.96 compared to GPT-3.5.

In the following section, we further validate our claims regarding NACo's effectiveness in QG assessment through human evaluation.

4.3 Human Evaluation

We recruit volunteer annotators, all fluent English speakers, to evaluate both model-generated questions and human-written questions. We use the 48 test instances from HotpotQA. For each instance, annotators evaluate four questions: RefQ, GPT-3.5 (zero-shot), BART-base, and AnnoQ. To ensure impartiality, we anonymize and shuffle these candidate questions before presenting them to the annotators. They are tasked with rating each question based on naturalness, answerability, and complexity, using a 3-point scale for each criterion (with 2 as acceptance and 0 as rejection). Additionally, we sum the individual scores together to calculate a combined score that reflects the question's overall quality. We obtain three annotations per question and use the average of these as the standard for human judgment. The Pearson correlation coefficient between ratings given by our annotators are 0.67. The rating rubric is available in Appendix A.4.

We calculate the Pearson r, Spearman ρ, and Kendall τ correlation coefficients to measure the agreement between all metrics, including NACo, and human judgment, as reported in Table 3. This comparison considers correlation with both indi-

Metric	Naturalness			Answerability			Complexity			**Overall**		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
B	-0.26	-0.24	-0.20	-0.19	-0.17	-0.15	0.00	0.02	0.02	-0.13	-0.07	-0.06
B-RT	0.11	0.09	0.07	0.16	0.14	0.11	0.33	0.33	0.25	0.30	0.28	0.21
R-L	0.09	0.04	0.03	0.11	0.06	0.05	0.30	0.27	0.22	0.25	0.19	0.15
BSc	0.19	0.18	0.15	0.21	0.22	0.18	0.39	**0.41**	**0.33**	0.37	0.38	0.28
Q-B	0.09	0.05	0.04	0.13	0.08	0.07	0.33	0.34	0.26	0.28	0.25	0.18
QA-S	0.10	0.09	0.06	0.04	0.02	0.02	0.06	0.02	0.01	0.03	0.02	0.01
R-Q	0.36	**0.25**	**0.19**	0.69	0.54	0.43	0.37	0.29	0.22	0.62	0.43	0.32
NACo	**0.40**	**0.25**	**0.19**	**0.70**	**0.59**	**0.49**	**0.44**	0.34	0.27	**0.68**	**0.49**	**0.38**

Table 3: Correlation between human assessments and automated evaluation metrics as indicated by Pearson r, Spearman ρ, and Kendall τ correlation coefficients. The highest and second-highest scores are highlighted with bold and underline markers, respectively.

vidual criteria and the overall question quality.

Table 3 reveals that NACo demonstrates the highest correlation with human evaluation for individual criteria in 7 out of 9 scores, and it ranks second-highest in the remaining 2 scores. Notably, NACo exhibits the strongest agreement with human judgment concerning the overall quality of questions across all correlation metrics, outperforming the state-of-the-art RQGUE metric by 6% in both r, ρ, and τ correlations.

5 Related Work

Evaluation Metrics for Question Generation: The evaluation of QG models relies heavily on automatic evaluation, where examples that score well and poorly on the automatic metrics serve as indicators of potential issues that new approaches should address. The most commonly used metrics are reference-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020). Based on correlation with human judgment, there have been studies attempting to challenge the effectiveness of these reference-based metrics and propose reference-free evaluation mechanism for QG (Nema and Khapra, 2018; Ji et al., 2022; Mohammadshahi et al., 2023). Our study, however, questions the competency of reference-based metrics by replicating the data collection process of benchmarks and introducing new references.

LLMs as evaluators for NLG tasks: A growing research interest revolves around the use of large language models (LLMs) for evaluating quality of generated texts (Liu et al., 2023; Wang et al., 2023; Lin and Chen, 2023; Chiang and Lee, 2023). Investigating GPT-3 and its variances' evaluation capacity on story generation and adversarial attack tasks, Chiang and Lee found that when given the same instructions as human annotators, LLMs show positive correlation with human judgment. Lin and Chen and Liu et al. obtained similar observations for dialogue generation and text summarization tasks. Due to the recent nature of this research direction, no other work has performed a comprehensive study on the use of LLMs as evaluators for the question generation task.

6 Conclusion

In this work, we questioned the competency of reference-based metrics in providing an accurate assessment for question generation. We replicated the data collection process used for benchmark datasets, which allowed us to gather candidate questions qualified as new references. Our analysis highlights the shortcomings of reference-based metrics in differentiating new references from less desirable candidates, assigning significantly lower scores to the former. Even the recently introduced reference-free metric, RQUGE, face difficulties in this regard. To address these challenges, we introduce NACo, a multi-dimensional, reference-free metric designed to bridge the gap between automated evaluation and human judgment in question generation. Our experimental results showcase that NACo, leveraging the Chain-of-Thought capabilities of Large Language Models for question answering, not only meets the expectations for quantitative QG metrics but also achieves state-of-the-art alignment with human evaluation.

Limitations

A limitation of our work speaks to the required access to a reasonable number of references to assess domain-specific or dataset-specific complexity. Future works can investigate how to account for expected complexity in scenarios where references are limited and difficult to collect. Moreover, NACo, like other reference-free metrics for QG, is subject to the performance of the underlying QA model. Specifically, the constraints of GPT-3.5 in answering complex, multi-hop questions might have limited NACo's ability to evaluate valid references closer to the upperbound. Future directions should explore evaluation frameworks that are robust to variations in QA model performance.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4645–4654, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3121–3131, Hong Kong, China. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Tianbo Ji, Chenyang Lyu, Gareth Jones, Liting Zhou, and Yvette Graham. 2022. Qascore—an unsupervised unreferenced metric for the question generation evaluation. *Entropy*, 24(11):1514.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, WWW '20, page 2032–2043, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liuyin Wang, Zihan Xu, Zibo Lin, Haitao Zheng, and Ying Shen. 2020. Answer-driven deep question generation based on reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5159–5170, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiaojing Yu and Anxiao Jiang. 2021. Expanding, retrieving and infilling: Diversifying cross-domain question generation with flexible templates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3202–3212, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 What makes a good question?

After the best model for question generation has been developed, it often goes through a round of human evaluation to assess the quality the generated questions. The human evaluation stage often looks at the following aspects of the generated question:

Naturalness (Wang et al., 2020; Bi et al., 2021) addresses essential linguistic elements of a question, such as whether the question is free from grammar mistakes (Ushio et al., 2022), or how clear and fluent the question sounds (Pan et al., 2020; Laban et al., 2022).

Answerability measures how well the question is grounded to the input context and answer. In this sense, a good question should be relevant to the input context (Pan et al., 2020; Wang et al., 2020), and in the answer-aware setting, should have a reasoning path that leads to the given answer (Ushio et al., 2022; Ji et al., 2022; Nema and Khapra, 2018; Mohammadshahi et al., 2023).

Complexity (Wang et al., 2020; Bi et al., 2021) speaks to the reasoning path taken to answer the question. The higher number of reasoning steps

Criterion 1: Naturalness
Natural Question: What 1996 book was written by the founder of Media Matter for America?
Unnatural Question: In 1996, what book did the founder of Media Matter for America, he write it?

Criterion 2: Answerability
Answerable Question: Which Australian actress stars in the black comedy sequel of "Forgetting Sarah Marshall"? (*given answer: Rose Byrne, actual answer: Rose Byrne*)
Unanswerable Question: Which black comedy sequel to "Forgetting Sarah Marshall" starred an Australian actress? (*given answer: Rose Byrne, actual answer: Get Him to the Greek*)

Criterion 3: Complexity
Passage and Answer: Although the two displayed great respect and admiration for each other, their friendship was uneasy and had some qualities of a love-hate relationship. Harold C. Schonberg believes that Chopin displayed a "tinge of jealousy and spite" [...] **Liszt** was the dedicatee of Chopin's Op. 10 Études, and his performance of them prompted the composer to write to Hiller, "I should like to rob him of the way he plays my studies."
Less complex question: Who did Chopin dedicate the Op. 10 Études to?
1. *The passage states that Liszt was the dedicatee of Chopin's Op. 10 Études.*
2. *Answer: Liszt*

More complex question: With whom was Chopin said to have a love-hate relationship?
1. *The passage mentions that Chopin had a love-hate relationship with someone.*
2. *The passage provides information about Chopin's relationship with Liszt, including admiration and annoyance.*
3. *Answer: Liszt*

Figure 4: Examples for each criterion addressed by our metric: Naturalness, Answerability, and Complexity.

needed, the more complex the question. It should be noted that higher complexity does not necessarily indicate better quality in a question. This quality rather depends on the nature of the dataset.

Figure 4 illustrates the gap between current automatic QG metrics and human evaluated metrics, where two questions using similar words can have opposite qualities. This gap can be explained by the fact that existing automatic metrics do not directly address any of the criteria that human annotation often looks for in a question. To address this challenge, our metric integrates the human perspective of a "good" question: naturalness, answerability, and complexity, into the evaluation pipeline.

A.2 Prompt for CoT-QA

You will be given [one/two] context passage(s) and a sentence. If the sentece is a question, your task is to output a text span from the context passage to answer the question. Your answer should NOT be complete sentences.
Instructions:

1. Let's read the passage first and then read the sentence. Consider:
   (a) Is the sentence a question? If yes, what information indicates that it is a question? If not, output 'not a question' and stop generation.
   (b) If it is a question, considers if the question is unclear, or has grammar errors. If so, output 'Question unnatural'.
2. Now find the answer to the question. Speak out loud your detailed reasoning.
3. Highlight your answer between two <ans> tokens.

Format you response as follows:

1. Your response to 1a and 1b
2. Step by step reasoning:
   (a) Step 1...
   (b) Step 2...
3. Answer: <ans> [answer text] <ans>

Context Passage 1: [Context Passage 1]
Context Passage 2: [Context Passage 2 if available]
Sentence: [Question to be evaluated]
Response:

A.3 Additional results for reference-based metrics

We report the performance of QG competitors across reference-based metrics, using questions created by our annotators (AnnoQ) as references, in Table 4 and 5.

A.4 Instructions for Human Evaluation

In this survey, you will be annotating 10 examples. For each example, you are given 2 passages that share some common information. A text span from one of the two passages will be bolded, italicized, and highlighted in blue. Your task is to rate 4 candidate questions on a scale of 0-2 for each of the following aspects:
Fluency: Does the question make at least one of the following errors: (1) grammar mistakes, (2) unclear objectives, or (3) not a question?

QG Competitor	Ref-based metrics					Ref-free metrics		Our metric			
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	Na.	Ans.	Cmp.	NACo
LM-generated											
BART-base	17.92	-0.30	43.29	92.03	37.24	**-0.37**	4.62	99.60	66.75	69.15	73.30
GPT-3.5 (few-shot)	21.04	-0.19	44.61	92.44	38.49	**-0.37**	4.56	**100.00**	65.23	69.63	73.67
BART-clue-RefQ	13.88	-0.37	40.43	91.74	33.2	-0.38	4.56	99.73	63.94	69.24	69.97
Human-validated											
RefQ	12.78	-0.31	37.83	91.52	31.32	-0.38	**4.89**	99.87	**68.67**	68.93	**75.09**
AnnoQ	100.00	100.00	100.00	100.00	100.00	**-0.37**	4.71	99.73	67.70	70.33	74.01
AnnoQ-clue-RefQ	**42.42**	**0.06**	**62.54**	**94.62**	**55.73**	-0.38	4.76	99.87	66.34	**70.36**	74.21

Table 4: **SQuAD** - Performance of different QG methods on NACo and other existing metrics. The evaluation uses our annotated questions (AnnoQ) as references, with GPT-3.5 as the underlying QA system. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.

QG Competitor	Ref-based metrics					Ref-free metrics		Our metric			
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	Na.	Ans.	Cmp.	NACo
LM-generated											
BART-base	12.26	-0.87	30.11	87.05	24.26	-0.28	2.78	64.58	31.00	68.56	41.25
GPT-3.5 (zero-shot)	13.36	-0.84	33.34	88.94	20.46	-0.28	4.25	**100.00**	71.55	70.72	73.60
GPT-3.5 (few-shot)	14.19	-0.69	34.54	89.45	23.94	**-0.27**	4.21	99.31	67.31	72.64	71.09
BART-clue-RefQ	12.78	-0.87	34.02	88.78	27.02	-0.28	3.31	90.28	46.84	79.41	63.17
Human-validated											
RefQ	13.64	-0.69	30.94	89.22	26.94	**-0.27**	4.21	99.31	59.62	**83.02**	74.33
AnnoQ	100.00	100.00	100.00	100.00	100.00	**-0.27**	4.32	**100.00**	**76.11**	82.05	**83.43**
AnnoQ-clue-RefQ	**33.66**	**-0.35**	**50.60**	**93.08**	**43.93**	**-0.27**	**4.44**	**100.00**	71.46	81.46	81.60

Table 5: **HotpotQA** - Performance of different QG methods on NACo and other existing metrics. The evaluation uses our annotated questions (AnnoQ) as references, with GPT-3.5 as the underlying QA system. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.

- If the question does not make any errors, give a 2 for this criterion
- If the question makes 1 of the above errors, give a 1 for this criterion
- If the question makes at least 2 of the above errors, give a 0 for this criterion

Answerability: Try answering each question yourself. An acceptable question should be relevant to the context passages and has a reasoning path that leads to the given answer highlighted in blue.

- If the answer to the candidate question is exactly the text highlighted in blue, give a 2 for this criterion
- If the answer to the candidate question contains some but not all parts of the text highlighted in blue, or contains all parts of the text highlighted in blue but with extra information, give a 1 for this criterion
- If the answer to the candidate question does not match the text highlighted in blue at all, give a 0 for this criterion.

Complexity: Try answering each question yourself. Does the question require reasoning over both passages? An acceptable question should use information from both passages, not just one.

- If you need to read both passages to answer the question, give a 2 for this criterion
- If you need to read only one passage to answer the question, give a 1 for this criterion.
- If you do not need any of the passages to answer the question, give a 0 for this criterion.

A.5 Experiment Details

Our **BART-base QG models** are initialized from checkpoint `facebook/bart-base`, which has 139M parameters, and further finetuned on the specific QG dataset (SQuAD or HotpotQA). All models are implemented with Hugging Face Transformers 4.20. We add two special tokens: (1) `<ans>` -

used to highlight the answer span in the context input, and (2) <clue> - used to highlight the clue words in the context input (for BART-clue-RefQ). The model is finetuned with a batch size of 128, a learning rate of $1e-4$, a maximum input length of 512, and a maximum output length of 32. The best model is selected based on the lowest validation loss.

For **GPT3.5** (gpt3.5-turbo) **QG and CoT-QA**, we employ the default hyperparameters with a temperature of 0.7 and top_p of 1. In our CoT-QA experiments on SQuAD, given the larger sample size of 750 and cost constraints, we conducted a single run. For our CoT-QA experiments on HotpotQA, we carried out 3 runs on the 50-examples sample and report the average scores from the three responses.

Implementations of existing metrics: We use the implementation of Hugging Face evaluate[1] package for BLEU (bleu), ROUGE (rouge), BLEURT (bleurt), BERTScore (bertscore), and RQUGE (rquge). We use the code released by the original papers to obtain implementation of QAScore[2] and Q-BLEU[3].

[1] https://huggingface.co/docs/evaluate/en/index
[2] https://github.com/TianboJi/QAScore/tree/main
[3] https://github.com/PrekshaNema25/Answerability-Metric