Reference-based Metrics Disprove Themselves in Question Generation

Bang Nguyen
University of Notre Dame
bnguyen5@nd.edu

Mengxia Yu
University of Notre Dame
myu2@nd.edu

Yun Huang
University of Illinois Urbana-Champaign
yunhuang@illinois.edu

Meng Jiang
University of Notre Dame
mjiang2@nd.edu

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 replicate the annotation process and collect another reference. A good metric is 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 reference-based metrics, developed a QA model, and defined a new metric named "answerability" or RQUGE. Though they showed a 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; Oh et al., 2023; Tang et al., 2023). However, QG benchmarks often contain only 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. Besides the new references, we 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. Fig. 1 highlights the incompetency of current QG metrics in distinguishing the new ref-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13651–13666
November 12-16, 2024 ©2024 Association for Computational Linguistics

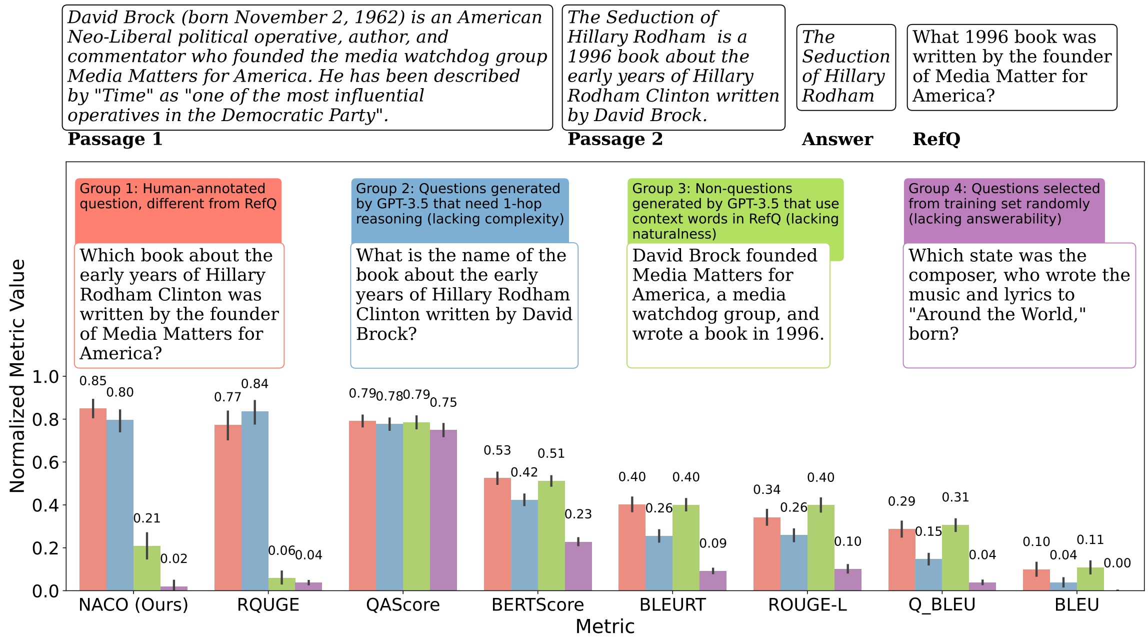


Figure 1: Normalized value of different evaluation metrics for four types of candidate questions against the same reference (RefQ) in the HotpotQA dataset (Yang et al., 2018). Ideally, metrics should score Group 1 highest. Current QG metrics, except for NACo (ours) and RQUGE, primarily recognize random questions (Group 4) but fail to differentiate between Groups 1 and 3 (note the red and green bars). RQUGE, successfully identifies groups violating naturalness (Group 3) and answerability (Group 4), assigns a higher score for Group 2, which lacks complexity, than for Group 1. Our metric, shown in the leftmost bar group, prioritizing essential criteria of a question, can effectively distinguish all four groups of candidates while maintaining the highest rating for the valid questions.

erence (a valid question; 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 neglect: (1) **N**aturalness: *how natural the question sounds* (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 address the challenges of evaluating question quality without access to a diverse set of references.

To overcome the limitation of the answerability measure in RQUGE (Mohammadshahi et al., 2023) 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 leftmost group of bars in Fig. 1 shows 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 question quality.

- 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.

We release the collected data and code implementation of NACo to facilitate future works. [1]

2 Failure of Reference-based QG Metrics

2.1 Study Design & Data Collection

Previous studies questioning 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 QG datasets. Specifically, we replicate the data collection procedure of the datasets 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 metrics. In addition, we extended our collection procedure 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. Fig. 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 creating their first question.

We perform the additional annotation on two popular QG benchmarks: (1) 748 test examples of SQuAD (Rajpurkar et al., 2016), and (2) 96 test examples of HotpotQA (Yang et al., 2018). To illustrate the application of our study, we collect another QG dataset in the educational domain, specifically from the TED-Ed learning platform[2]. We further annotate 43 questions from this new dataset. More details about data collection and annotation for Ted-EdQA are provided in Appx. A.5.

For the HotpotQA sample, we also collect 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 select one from the original two passages that contains the answer span. We then ask GPT-3.5 to generate a question based on this single passage. We review the questions for grammar, clarity, relevance to the passage, independence from external knowledge, and a logical path to the answer.

Group 3: Non-questions that use the same words as the reference: For this group of questions, we ask 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 add a constraint such that the generated sentence cannot be in the form of a question. We then manually go through the generated sentences to ensure that no hallucinations were in place. In this sense, we produce 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.

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 Fig. 1, the answer to this candidate question is *Robin McLau-*

[1] https://github.com/bnguyen5/naco

[2] https://ed.ted.com/

rin Williams, which is completely irrelevant to the given answer *The Seduction of Hillary Rodham*. In this sense, this group of candidate questions violates the answerability aspect of an ideal candidate.

2.2 Results

Fig. 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 scores. For instance, the average ROUGE-L score for Group 4 is 0.10, compared to 0.34 for Group 1, 0.26 for Group 2, and 0.40 for Group 3, with a minimum difference of 16% from the scores of the other groups.

Fig. 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 (green bar) on average 6% higher than a new reference produced by our annotators (red bar), and 14% higher than a perfectly answerable question requiring less reasoning than the reference (blue bar). This observation 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 (0.84) 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 0.77.

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

Based on extensive review of the human evaluation procedure in QG literature, detailed in Appx. A.1, we identify three essential criteria of a question: **N**aturalness, **A**nswerability, and **C**omplexity. We propose NACo, which leverages prompting and Chain-of-Thought (CoT) reasoning (Wei et al., 2022) to obtain a score for each criterion. Specifically, given the relevant context passage(s) and a question, we instruct an LLM as follows:

- The LLM first reads over the context passage(s) and the question. The LLM checks 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 provides the final answer to the question. We define the **answerability** of the question 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 stan-

QG Competitor	Ref-based metrics					NACo
	B	B-RT	R-L	BSc	Q-B	
LM-generated						
BART-base	19.53	-0.28	44.79	92.13	36.94	73.30
GPT-3.5 (few-shot)	18.06	-0.23	43.58	92.18	36.48	73.67
BART-clue-RefQ	**31.91**	**0.07**	**59.92**	**94.37**	**52.33**	69.97
Human-validated						
RefQ	100.00	1.00	100.00	100.00	100.00	**75.09**
AnnoQ	12.78	-0.31	37.83	91.52	31.32	74.01
AnnoQ-clue-RefQ	27.43	0.04	53.62	93.85	46.89	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 LLM. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.

dards for the domain and dataset. For example, in the HotpotQA dataset, questions that require multi-hop reasoning might be preferred over simpler, single-hop questions. This preference may not hold in 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. In our experiments, we use 750 examples from the training set of SQuAD and HotpotQA to compute the expected complexity for each dataset. 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. We provide additional details on how c_{cand} is computed and integrated into the final score in Appx. A.3

4 Experiments

4.1 Experimental Setup

Question generation competitors: 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 also produce another version of BART-base, **BART-clue-RefQ**, which highlight the ground-truth clues used by reference questions (RefQ) in the context given as input to BART-base (detailed in A.6). 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. 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.

Baselines: We compare the evaluation capacity of NACo with five reference-based metrics, including BLEU-4 (B) (Papineni et al., 2002), BLEURT (B-RT) (Sellam et al., 2020), ROUGE-L (R-L) (Lin, 2004), BERTScore (BSc) (Zhang et al., 2019), and Q-BLEU (Q-B) (Nema and Khapra, 2018), and two reference-free metrics, QAScore (QA-S) (Ji et al., 2022), and RQUGE (R-Q)(Mohammadshahi et al., 2023). Tang et al. proposes using LLM to diversify the limited references in benchmarks, demonstrating an improvement in the correlation between reference-based metrics and human judgment. We replicate this approach and report the evaluation performance of the five reference-based metrics both when only the original reference is used and when adding the diversified references.

NACo implementation: We provide the CoT prompt used in our experiments in Appx. A.2. We experimented with five underlying LLMs: Llama3-8B, Mixtral-8x7B, Claude3-Haiku, GPT3.5-turbo, and GPT4o.

Human Evaluation: We recruit volunteer annotators, all fluent English speakers, to evaluate both model-generated questions and human-written questions, using 96 test examples from HotpotQA. For each example, annotators evaluate four questions: RefQ, GPT-3.5 (zero-shot), BART-base, and AnnoQ, displayed in randomized and anonymized order. Evaluators rate each question based on naturalness, answerability, and complexity, using a 3-point scale for each criterion. Additionally, we sum the individual scores 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

Metric	Naturalness			Answerability			Complexity			**Overall**		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Ref-based metric												
B	0.08	-0.06	-0.04	0.16	0.06	0.05	0.40	0.38	0.31	0.25	0.20	0.16
w/ *DivRef*	0.09	-0.01	-0.01	0.18	0.10	0.08	0.43	0.42	0.33	0.27	0.25	0.19
B-RT	0.12	0.03	0.03	0.23	0.19	0.14	0.49	0.50	0.37	0.33	0.35	0.24
w/ *DivRef*	0.14	0.05	0.04	0.26	0.22	0.16	0.51	0.52	0.39	0.36	0.38	0.26
R-L	0.07	-0.03	-0.03	0.18	0.10	0.07	0.43	0.42	0.31	0.27	0.24	0.17
w/ *DivRef*	0.13	0.06	0.04	0.24	0.19	0.15	0.48	0.48	0.37	0.34	0.34	0.24
BSc	0.18	0.09	0.06	0.27	0.21	0.16	0.52	0.54	0.41	0.38	0.39	0.27
w/ *DivRef*	0.23	0.14	0.10	0.33	0.29	0.21	**0.56**	**0.57**	**0.43**	0.44	0.45	0.31
Q-B	0.07	-0.05	-0.04	0.17	0.09	0.07	0.45	0.44	0.32	0.27	0.26	0.17
w/ *DivRef*	0.07	-0.04	-0.03	0.16	0.09	0.07	0.45	0.44	0.32	0.27	0.26	0.18
Ref-free metric												
QA-S	-0.01	0.00	0.00	0.01	-0.01	-0.01	0.04	0.03	0.02	0.02	0.02	0.02
R-Q	0.38	0.26	0.20	0.67	0.51	0.39	0.33	0.22	0.16	0.57	0.38	0.27
NACo (Ours)												
Llama3-8B	0.49	0.32	0.25	0.66	0.50	0.39	0.43	0.36	0.21	0.64	0.42	0.30
Mixtral-8x7B	0.36	0.30	0.23	0.54	0.52	0.40	0.36	0.23	0.17	0.56	0.40	0.28
Claude-Haiku	0.53	0.30	0.23	0.71	0.51	0.40	0.50	0.35	0.27	0.71	0.47	0.35
GPT3.5	0.47	0.30	0.23	0.70	0.53	0.41	0.49	0.35	0.27	0.68	0.48	0.35
GPT4	**0.64**	**0.36**	**0.28**	**0.78**	**0.59**	**0.47**	0.55	0.35	0.27	**0.80**	**0.52**	**0.39**

Table 2: Correlation between human assessments and automated evaluation metrics as indicated by Pearson r, Spearman ρ, and Kendall τ correlation coefficients. For reference-based metrics, we report the metric's correlation with human judgment both when only the original reference is used and when adding the diversified references (w/ *DivRef*). The highest and second-highest scores are highlighted with bold and underline markers, respectively. Shaded regions indicate an improvement compared to the current state-of-the-art metric for that respective column.

ratings given by our annotators is 0.67. The rating rubric is available in Appx. A.4.1.

4.2 Results

Failure of reference-based metrics: We report QG competitors' performance on various metrics, including NACo, using RefQ as the reference in Tbl. 1 for SQuAD. Even though RefQ, AnnoQ, and AnnoQ-clue-RefQ are all qualified as valid questions, reference-based metrics rate them with significant differences. In the SQuAD dataset, BLEU scores for RefQ, AnnoQ, and AnnoQ-clue-RefQ are 100, 12.78, and 27.43, respectively (Tbl. 1). However, NACo rates these three groups of questions similarly, with RefQ, AnnoQ, and AnnoQ-clue-RefQ scoring 75.09, 74.01, and 74.21, respectively (Tbl. 1). Similar patterns are observed in the HotpotQA and TedEdQA datasets, as detailed in Tbl. 7 and Tbl. 8.

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 to BLEU-4, reported in Tbl. 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.

Effectiveness of NACo: Referring to our analysis of four groups of candidate questions for HotpotQA in Fig. 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 satisfy all criteria for HotpotQA questions, achieve the highest average NACo score of 0.85. They are followed by the questions in Group 2, lacking in complexity, with a score of 0.80; Group 3, lacking in naturalness, with a score of 0.21; and Group 4, lacking in answerability, with a score of 0.02.

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 Tbl. 2. This

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.

comparison considers correlation with both individual criteria and the overall question quality. Tbl. 2 reveals that NACo demonstrates the highest correlation with human evaluation for individual criteria in 9 out of 12 scores. Notably, NACo exhibits the strongest agreement with human judgment concerning the overall quality of questions across all correlation metrics. This observation is consistent across different underlying LLMs.

4.3 Analysis

QG Competitor	B	R-Q	NACo	Human
LM-generated				
BART-base	14.57	2.90	42.65	2.80
GPT-3.5 (zero-shot)	9.46	4.18	74.99	4.60
Human-validated				
RefQ	100.00	4.12	75.32	5.14
AnnoQ	14.80	4.21	84.97	5.45

Table 3: **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 for NACo.

NACo vs. Reference-based Metrics: Tbl. 3 indicates that reference-based metrics rate BART-base questions slightly lower than AnnoQ (by 0.23% according to BLEU), whereas NACo shows a much larger gap (42.32%). 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. This observation is validated by our human evaluators, detailed in A.4.2. Fig. 2 provides a case study where BERTScore, the reference-based metric most aligned with human judgment (Mohammadshahi et al., 2023), favored BART-base generation over the human annotated question, even though the former was not formatted as a question. 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, but fail to form 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).

NACo vs. Existing Reference-free Metrics: Tbl. 3 also reveals that the new reference-free metric for QG, RQGUE, rates GPT-3.5 generated questions—whether in zero-shot or few-shot modes—comparably to the original reference question (RefQ). A manual review showed that GPT-3.5 typically utilizes only one of two context passages for creating a multi-hop question, as illustrated in Fig. 3. Again, human evaluation verifies our observations, detailed in Appx. A.4.2. In the case study, GPT-3.5 exclusively used context words from Passage 1, making access to Passage 2 unnecessary for answering the question. Meanwhile, RefQ incorporates context words from both passages and requires reasoning across both for an answer. RQUGE overlooks this aspect and assigns a higher score for the GPT-3.5 question than for RefQ (94.52 and 93.17, respectively). Addressing this gap, NACo acknowledges the answerability and naturalness of the GPT-3.5 question, but penal-

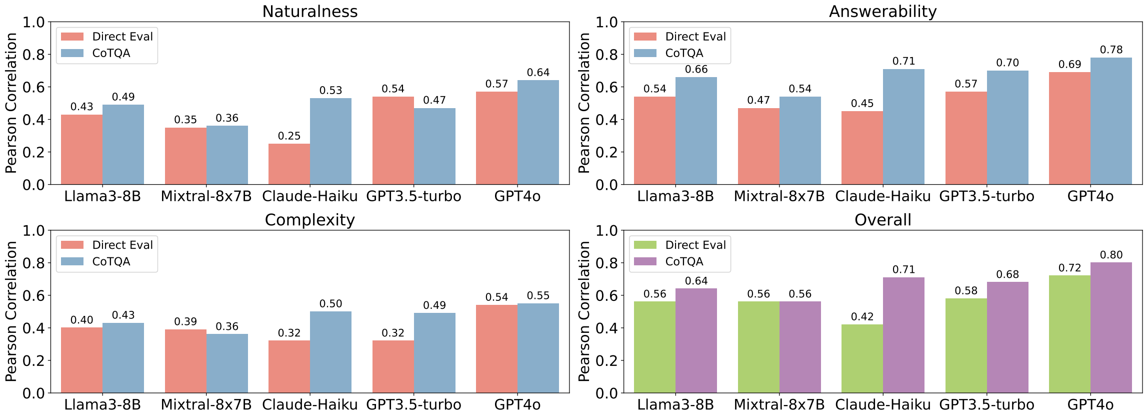


Figure 4: Correlation with human judgement - Comparing CoT-QA (NACo) with Direct Evaluation

izes 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.

Human Preference Study: As our human evaluation study assesses candidate questions based on the criteria that NACo measures, we conduct a human preference study to ensure fair comparisons between NACo and existing metrics. In this study, we further compare NACo with the top two baselines: RQUGE and BERTScore. We sample 20 pairs of questions where NACo and each baseline disagree in their scoring. Three human evaluators make their preferences for each pair, achieving a Cohen's Kappa of 0.87, indicating strong agreement. Using human preference as the reference, NACo wins against RQUGE in 15 out of 20 cases (75%) and BERTScore in 12 out of 20 cases (60%).

NACo (CoT-QA) vs. LLM Direct Evaluation: Large language models (LLMs) are increasingly utilized as proxies for human evaluators. Previous studies have suggested that when receiving CoT instructions typically given to human evaluators, LLMs can assess generated texts in a way that are highly aligned with human judgement (Liu et al., 2023). We refer to this approach of using LLM evaluators as Direct Evaluation (DirectEval). We examine the effectiveness of CoT-QA, used by NACo, against DirectEval, which provides the LLMs the same human evaluation instructions in Appx. A.4.1. Fig. 4 presents the Pearson correlation coefficients, comparing the performance of CoT-QA (NACo) with DirectEval across individual criteria and overall question quality. The results indicate a higher alignment with human judgment when employing CoT-QA for each respective LLM. Notably, adopting CoT-QA instead of DirectEval significantly boosts the performance of Claude-Haiku, improving the alignment with human judgment of overall question quality from 0.42 to 0.71. This improvement is comparable to the performance achieved using GPT4o (0.72 in DirectEval setting, 0.80 in CoTQA setting), while being 12 times more cost effective.

We also investigate whether DirectEval and NACo will benefit from the addition of a reference question during evaluation. For DirectEval, we provide the reference question at the end of the instruction. For NACo, we use the reference question's complexity (obtained from CoT-QA) as the expected complexity. In our experiment, we use the RefQ group as the reference, and evaluate the other three groups of candidate questions (BART-base, GPT3.5, and AnnoQ). Table 4 details the correlation between human judgement and the use of LLM evaluators with and without references.

Method	GPT3.5	Claude-Haiku
DirectEval	0.63	0.48
DirectEval + RefQ	0.74	0.52
NACo	0.72	0.74
NACo + RefQ	0.69	0.75

Table 4: Pearson correlation between human assessments and LLM evaluators, with and wihtout references (RefQ).

It can be seen that while references benefit DirectEval to some extent, they do not consistently improve it to the level of NACo. Specifically, DirectEval + RefQ with GPT3.5 shows on-par performance with NACo (0.74 v.s. 0.72), while DirectEval + RefQ with Claude-Haiku underperforms NACo by a large margin. Furthermore, adding references on top of NACo does not further improve its performance. These results suggest that a single reference does not provide orthogonal benefits to NACo, which aligns with our findings regarding the limitations of reference-based metrics. NACo, as a reference-free metric, provides comprehensive and robust QG evaluations, without suffering from the bias of a single reference and the expensive reference collection process.

5 Related Work

Evaluation Metrics for Question Generation: The evaluation of QG models commonly used 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, on the other hand, questions the competency of reference-based metrics by replicating the data collection process of benchmarks and introducing new references. Other works have taken a similar approach, designing and collecting different groups of candidates to investigate reference-based metrics in machine translations (Amrhein et al., 2022; Karpinska et al., 2022) and question answering (Bulian et al., 2022). However, QG poses unique challenges to the evaluation of question quality, considering aspects such as complexity and answerability, and therefore call for a study like ours.

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, gathering candidate questions qualified as new references. Our analysis highlights the shortcomings of reference-based metrics in differentiating new references from flawed 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 bridging 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. We provide a case study to illustrate this issue in Appx A.7. Future directions should explore evaluation frameworks that are robust to variations in QA model performance.

Acknowledgements

This work was supported by NSF IIS-2119531, IIS-2137396, IIS-2142827, IIS-2234058, CCF-1901059, and ONR N00014-22-1-2507.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United

Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. 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.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Shinhyeok Oh, Hyojun Go, Hyeongdon Moon, Yunsung Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. 2023. Evaluation of question generation needs more references. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6358–6367, Toronto, Canada. 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*.

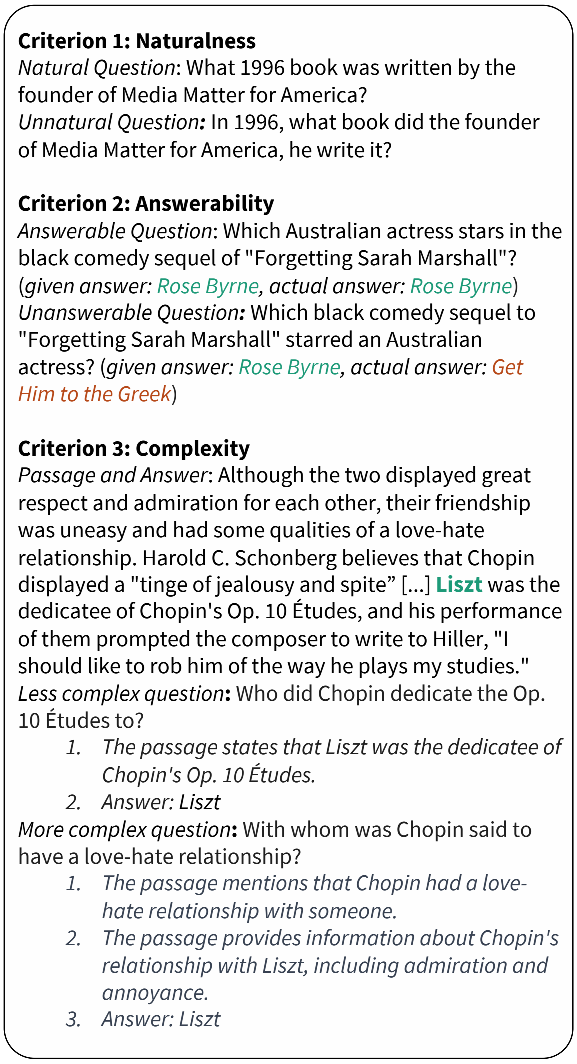
A Appendix

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



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

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 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.

Fig. 5 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 [reasoning step must be a single sentence with one clause]
   (b) Step 2 [reasoning step must be a single sentence with one clause]
   (c) ...
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 NACo Details

For each question, the Chain-of-Thought (CoT) QA prompt we provide to the LLM asks the model to output its question-answering process by steps, separated by newline characters. We post-process this formatted output to count the number of reasoning steps, referred to as the absolute complexity of the candidate question or $c_{\text{cand\_abs}}$.

To calculate the relative complexity of the candidate question with respect to the dataset, we first find the expected complexity associated with that dataset. Using a set of reference questions from the training set, we perform the same CoT QA process for each of these reference questions and obtain their absolute complexity. The expected complexity for the dataset is then the most common value (or mode) among the absolute complexities of the questions in this training sample, denoted as c_{expected}.

The final score regarding the complexity of the candidate question is the normalized value of the absolute difference between $c_{\text{cand\_abs}}$ and c_{expected}: $c_{\text{cand}} = 1 - \frac{|c_{\text{cand\_abs}} - c_{\text{expected}}|}{\max(c_{\text{cand\_abs}}, c_{\text{expected}})}$. By using $\max(c_{\text{cand\_abs}}, c_{\text{expected}})$, we ensure the range of c_{cand} is between 0 and 1. The final NACo score is then computed by taking a weighted sum of n_{cand} (binary, 0 or 1), a_{cand} (floating number between 0 and 1), and c_{cand} (floating number between 0 and 1). We used a weight of $\frac{1}{3}$ for each criterion score in our experiments, ensuring NACo's range to be between 0 and 1. In short: $\text{NACo} = \frac{1}{3}n_{\text{cand}} + \frac{1}{3}a_{\text{cand}} + \frac{1}{3}c_{\text{cand}}$.

A.4 Human Evaluation Details

A.4.1 Instructions

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?

- 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.4.2 Human Evaluation Results

QG Competitor	Nat. [0,2]	Ans. [0,2]	Cmp. [0,2]	Total [0,6]
BART-base	1.10	0.74	0.97	2.80
GPT-3.5	1.92	1.62	1.06	4.60
RefQ	1.70	1.68	1.75	5.14
AnnoQ	1.82	1.83	1.80	5.45

Table 5: Human Evaluation of QG Competitors on HotpotQA

A.5 TedEdQA Details

We collect 4246 multiple-choice questions from 1001 video lessons from TED-Ed[3]. Each data point comprises the transcript of the video lesson it is based on, the question stem, and the correct answer.. After excluding questions with answers such as *None of the above*, *All of the above*, *Both A and B*, etc., 3547 questions remain. We split the questions into three sets train, dev, and test, each with size of 3034, 259, and 254 respectively. We ensure that no questions from any set come from the same lecture as those in the other two sets.

From the test set, we select 43 questions (RefQ) derived from 12 video lessons for additional reference annotation. We follow similar procedures to the SQuAD and HotpotQA dataset that have annotators create two types of questions—one without clues and one with provided clues—based on

[3] https://ed.ted.com/

a given context and answer. However, the context presented to annotators differs: to formulate a reference-qualifying question, we provide them with the URL of the original lesson, the full transcript, and a specific context extracted from the transcript that is relevant to the answer. This extraction is conducted as the entire video transcript can be too long, potentially complicating the fine-tuning of models like BART. We obtain this extracted context by having GPT3.5-turbo label it from the full transcript and the original question RefQ. Specifically, we prompt the model: "*Given a lecture content and a multiple-choice quiz question, please extract the most relevant and concise context from the content that is best for creating the provided multiple-choice question. Ensure the extracted excerpt contains all the necessary information for creating the given quiz question*". This context is also used to fine-tune BART-base and to generate questions with GPT-3.5-turbo in a few-shot setting.

A.6 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.

Implementations of existing metrics: We use the implementation of Hugging Face `evaluate`[4] 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[5] and Q-BLEU[6].

For Div-Ref, which proposes diversifying references using LLM to improve reference-based metrics' alignment with human judgement, we use the same model settings as the authors (Tang et al., 2023). Specifically, we use GPT3.5-turbo with temperature set to 1 and top_p set to 0.9. We use 9/10 instructions proposed by Tang et al. 2023 to generate 9 new references from the original reference RefQ. (We did not use the remaining instruction because it asks the model to reorder sentences in a paragraph, while our text is only a question in the form of one sentence). When calculating the reference-based metric score across multiple references, we used the maximum aggregation.

LLM Details: We test 5 different LLMs for NACo. We interact with GPT3.5 (`gpt3.5-turbo`), GPT4o (`gpt-4o`)[7], Claude-Haiku (`claude-3-haiku-20240307`)[8], and Mixtral-8x7B (`open-mixtral-8x7b`)[9] through their official APIs. For Llama3-8B, we download the model via their Huggingface repository (`meta-llama/Meta-Llama-3-8B-Instruct`) and deploy it locally. All experiments with LLMs are used with their default hyperparameters. 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.

A.7 Error Analysis

We have noted that one of the limitations of NACo is the dependency of the QA models' performance. To further elaborate it, we provide the a case study in Fig. 6. The case study involves two context passages and the answer '*Teinosuke Kinugasa*.' We examine three candidate questions: a **GPT-3.5-generated** question (intended for 2-hop but resulting in 1-hop), the original HotpotQA reference (**RefQ**, 2-hop), and our newly collected reference (**AnnoQ-clue-RefQ**, 2-hop). The CoT-QA model employed by NACo (GPT3.5) correctly identifies 'Teinosuke Kinugasa' for both the GPT-3.5-generated question and AnnoQ-clue-RefQ but fails to do so for RefQ, responding with '*Not enough information provided to answer the question*.'

This failure with RefQ is attributed to its requirement for mathematical reasoning (subtracting birth year from death year), a task GPT-3.5 struggles with. Accordingly, NACo assigns the highest score to AnnoQ-clue-RefQ, fulfilling all three requirements, while penalizing the GPT-3.5 question for

[4] `https://huggingface.co/docs/evaluate/en/index`

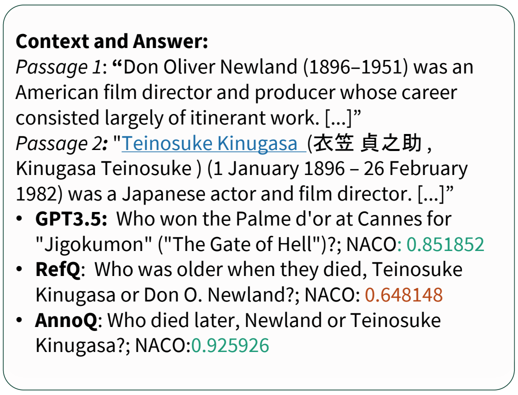
[5] `https://github.com/TianboJi/QAScore/tree/main`

[6] `https://github.com/PrekshaNema25/Answerability-Metric`

[7] `https://platform.openai.com/docs/overview`

[8] `https://www.anthropic.com/api`

[9] `https://docs.mistral.ai/api/`



Context and Answer:

Passage 1: **“**Don Oliver Newland (1896–1951) was an American film director and producer whose career consisted largely of itinerant work. [...]”

Passage 2: "Teinosuke Kinugasa (衣笠 貞之助 , Kinugasa Teinosuke) (1 January 1896 – 26 February 1982) was a Japanese actor and film director. [...]”

- **GPT3.5:** Who won the Palme d'or at Cannes for "Jigokumon" ("The Gate of Hell")?; NACO: 0.851852
- **RefQ**: Who was older when they died, Teinosuke Kinugasa or Don O. Newland?; NACO: 0.648148
- **AnnoQ**: Who died later, Newland or Teinosuke Kinugasa?; NACO:0.925926

Figure 6: NACo Error Analysis: Reliance on QA model capacity.

its simplicity and RefQ most severely (Answerability F1 score = 0) due to the mismatch between the CoT-QA answer and the provided answer.

A.8 Additional results for reference-based metrics

Tbl. 6 and 7 provides a more detailed version of 1 and 3.

Tbl. 8 illustrates the application of our study that disproves reference-based metrics and the proposed metric, NACo, in an educational setting using the TedEd-QA dataset. It can be seen that our observations regarding the failure of reference-based metrics and the effectiveness of NACo also holds for this dataset. Specifically, Refq, AnnoQ, and AnnoQ-clue-RefQ have significant gap when reference-based metrics are used to score them. NACo is able to score all these three human-validated candidates with similar scores and no worse than any machine-generated candidates.

QG Competitor	Ref-based metrics					Ref-free metrics		
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	NACo
LM-generated								
BART-base	19.53	-0.28	44.79	92.13	36.94	**-0.37**	4.62	73.30
GPT-3.5 (few-shot)	18.06	-0.23	43.58	92.18	36.48	**-0.37**	4.56	73.67
BART-clue-RefQ	**31.91**	**0.07**	**59.92**	**94.37**	**52.33**	-0.38	4.56	69.97
Human-validated								
RefQ	100.00	1.00	100.00	100.00	100.00	-0.38	**4.89**	**75.09**
AnnoQ	12.78	-0.31	37.83	91.52	31.32	**-0.37**	4.71	74.01
AnnoQ-clue-RefQ	27.43	0.04	53.62	93.85	46.89	-0.38	4.76	74.21

Table 6: **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 for NACo. 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			Human
	B	B-RT	R-L	BSc	Q-B	QA-S	R-Q	NACo	
LM-generated									
BART-base	14.57	-0.75	34.87	87.83	28.58	-0.28	2.90	42.26	2.80
GPT-3.5 (zero-shot)	9.46	-0.92	26.95	87.48	16.87	-0.28	4.18	74.99	4.60
GPT-3.5 (few-shot)	9.55	-0.86	27.48	87.71	17.33	-0.27	4.33	77.41	-
BART-clue-RefQ	**34.29**	-0.07	**61.40**	92.86	**53.94**	-0.28	3.51	62.40	-
Human-validated									
RefQ	100.00	1.00	100.00	100.00	100.00	-0.28	4.12	75.32	5.14
AnnoQ	14.80	-0.59	34.21	89.65	28.77	**-0.27**	4.21	**84.97**	5.45
AnnoQ-clue-RefQ	32.56	**-0.05**	53.21	**93.12**	52.85	**-0.27**	**4.26**	79.16	-

Table 7: **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 for NACo. 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	NACo
LM-generated								
BART-base	13.71	0.16	32.42	89.19	-	**-0.15**	3.74	79.29
GPT-3.5 (few-shot)	9.23	-0.28	28.77	88.6	-	-0.16	3.89	79.78
BART-clue-RefQ	13.63	0.01	40.66	90.09	-	**-0.15**	3.33	75.29
Human-validated								
RefQ	100.00	1.00	100.00	100.00	-	-0.16	3.99	82.85
AnnoQ	14.78	-0.43	34.16	89.61	-	-0.16	3.98	**84.67**
AnnoQ-clue-RefQ	**26.4**	**0.59**	**50.22**	**92.1**	-	-0.17	**4.23**	83.00

Table 8: **TedEdQA** - Performance of different QG methods using NACo and other existing metrics. The evaluation uses original TedEd questions (RefQ) as references, with GPT-3.5 as the underlying QA system for NACo. Some questions in this dataset are in the fill-in-the-blank form and do not contain question words like *what*, *where*, *etc*. which Q-BLEU heavily relies on for scoring (Nema and Khapra, 2018); thus, we do not report this metric for this dataset. The highest and second-highest scores (not including references for reference-based metrics) are highlighted with bold and underline markers, respectively.