RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question

Alireza Mohammadshahi[* 1,2,3] **Thomas Scialom**[1] **Majid Yazdani**[† 4]
Pouya Yanki[1] **Angela Fan**[1] **James Henderson**[2] **Marzieh Saeidi**[1]

[1] Meta AI   [2] IDIAP Research Institute   [3] EPFL   [4] BYJU's LAB

{alireza.mohammadshahi,james.henderson}@idiap.ch
{tscialom,pya,angelafan,marzieh}@meta.com
{majid.yazdani}@byjus.com

Abstract

Existing metrics for evaluating the quality of automatically generated questions such as BLEU, ROUGE, BERTScore, and BLEURT compare the reference and predicted questions, providing a high score when there is a considerable lexical overlap or semantic similarity between the candidate and the reference questions. This approach has two major shortcomings. First, we need expensive human-provided reference questions. Second, it penalises valid questions that may not have high lexical or semantic similarity to the reference questions. In this paper, we propose a new metric, RQUGE, based on the answerability of the candidate question given the context. The metric consists of a question-answering and a span scorer modules, using pre-trained models from existing literature, thus it can be used without any further training. We demonstrate that RQUGE has a higher correlation with human judgment without relying on the reference question. Additionally, RQUGE is shown to be more robust to several adversarial corruptions. Furthermore, we illustrate that we can significantly improve the performance of QA models on out-of-domain datasets by fine-tuning on synthetic data generated by a question generation model and re-ranked by RQUGE.[1]

1 Introduction

Given the context (e.g. paragraph), the goal of question generation (QG) is to generate questions with or without providing the answer spans. Automatic question generation can be used in several applications: improving the question answering (QA) task (Duan et al., 2017; Du and Cardie, 2018; Puri et al., 2020; Cheng et al., 2021), automatic assess-

Context: Today, **Warsaw** *has some of the best medical facilities in Poland and East-Central Europe.* The city is home to the *Children's Memorial Health Institute* (CMHI), …

Reference Question: Where are some of the best medical facilities in East-Central Europe located?

Question ① Where is the Children's Memorial Health Institute located?

❌**BLEU:**0.07 ❌**BERTScore:**0.44 ✅ **RQUGE:** 1.0

Question ② In which city of Poland, can the best medical equipments be found?

❌**BLEU:**0.12 ❌**BERTScore:**0.52 ✅ **RQUGE:** 1.0

Question ③ Where are some of the best medical facilities in West-Central Europe located?

❌**BLEU:**0.80 ❌**BERTScore:**0.96 ✅ **RQUGE:** 0.0

Figure 1: Normalised scores for different candidate questions. Metrics based on similarity to a reference question can penalise valid candidate questions, and compute a high score for unacceptable questions that are lexically similar to the reference. This can lead to the failure of reference-based metrics for valid questions, such as Q_1. Additionally, even paraphrases of the reference, like Q_2, may receive low scores. Furthermore, reference-based metrics may not detect small corruptions or variations in the reference, such as Q_3.

ments (Rebuffel et al., 2021; Lee et al., 2021), especially for the educational domain (Chen et al., 2018), and the evaluation of factual consistency in the text generation tasks (Scialom et al., 2019a, 2021; Fabbri et al., 2022).

Previous work (Hosking and Riedel, 2019; Scialom et al., 2019b; Zhang and Bansal, 2019; Laban et al., 2022) has shown that QG models can gen-

erate questions inconsistent with the corresponding context and the answer span. So, measuring the acceptability of candidate questions is a critical challenge. Human judgment is the most accurate method in natural language generation, but it is expensive, time-consuming, and not scalable. Consequently, several metrics e.g. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020) are proposed to automatically measure the quality of the generated text.

Specifically for the question generation task, previous work has utilised reference-based metrics e.g. BLEU, ROUGE, BERTScore, and BLEURT (Sellam et al., 2020; Ushio et al., 2023a,b, 2022) to evaluate the quality of the candidate question given the reference question. However, these methods highly depend on the diversity of the reference questions for a given answer span. Due to the huge cost of human annotations, existing QA/QG datasets mostly provide one reference question for the given context and answer, which results in wrongly penalising some valid questions. In Figure 1, the first candidate question (Q_1) is generated by paying attention to different evidence in the context, and Q_2 is a paraphrase of the reference, but both BLEU and BERTScore fail to assign high scores to them. Furthermore, reference-based metrics are not sensitive to very small corruptions of the reference questions, which makes the candidate question unacceptable (Q_3).

In this paper, we propose RQUGE , a **R**eference-free **QU**estion **G**eneration **E**valuation metric that can compute the quality of the candidate question without requiring a reference question. Given the corresponding context and answer span, our metric calculates the acceptability score by applying a general question-answering module, followed by a span scorer. The former module generates the answer span for the given candidate question, and the latter computes the semantic similarity of the predicted and gold answer spans. Our metric is extremely valuable in cases where the reference question is not well-formed [2] or there is one (or no) reference for a given context and answer span. We evaluate our metric on several datasets, including SQuAD (v1) (Rajpurkar et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019), and MS-MARCO (Bonifacio et al., 2021), and show

that it consistently has a better correlation with human judgment compared to previous QG metrics. We also integrate RQUGE into the decoding step by re-ranking candidate questions of each instance by our metric, leading to a better correlation with the human evaluation. Additionally, we demonstrate that RQUGE is more robust to adversaries than previous metrics with +13.1% relative improvement. Finally, we improve the performance of question answering models on an out-of-domain dataset by fine-tuning them on synthetic data generated by a question generation model, then re-ranked with RQUGE to choose the best candidate question for the given answer span, resulting in an +18.3% F1 and +22.2% EM relative improvement.

To sum up, our contributions are as follows:

- We propose RQUGE, an evaluation metric for measuring the quality of the automatically generated questions, without requiring access to any reference questions.

- We show that our metric has a significantly higher correlation with human judgment in terms of the acceptability of the candidate questions on SQuAD (v1), NQ, and MS-MARCO datasets. Also, re-ranking candidate questions with RQUGE leads to a better correlation with human judgment.

- We demonstrate that RQUGE metric is more robust compared to previous work on several adversarial strategies such as negation, entity swapping, gender reversing, or paraphrasing the reference questions.

- Finally, we illustrate that the performance of QA models significantly improves on the out-of-domain datasets by fine-tuning them on the synthetic data, created by applying a question generator model, then re-ranking with RQUGE metric.

2 Related Work

Previous work on automatic evaluation of Natural Language Generation (NLG) tasks have been categorized as follows:

Unsupervised Metrics. It contains the most commonly used metrics e.g. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), chrF (Popović, 2015), and METEOR (Denkowski and Lavie, 2010). These metrics calculate the correlation of reference

---

[2] For instance, many questions in the Natural Question dataset (Kwiatkowski et al., 2019) are not well-formed. Example: "in a deep mine temperatures increase with depth at the rate of"
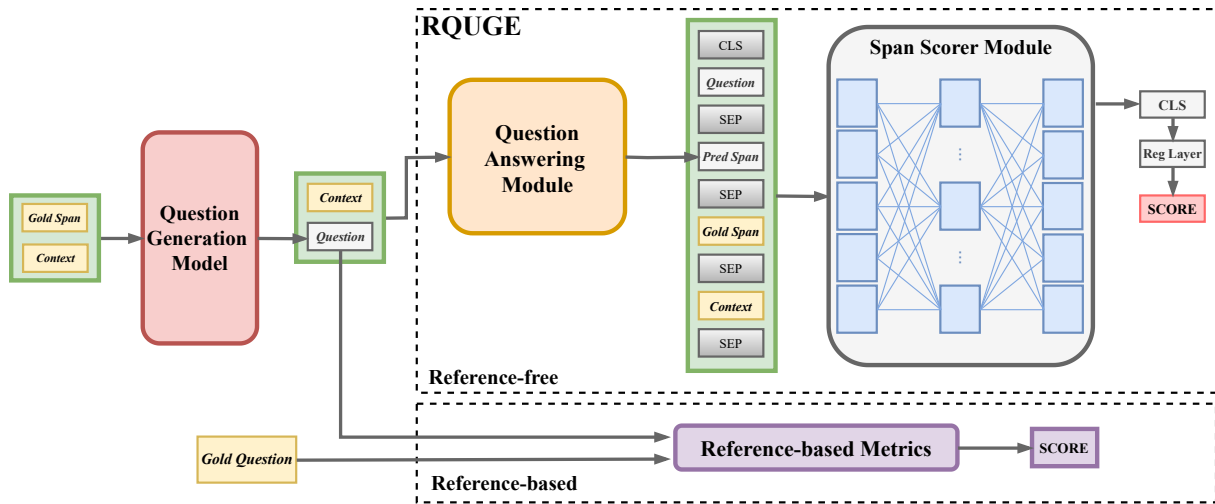
Figure 2: The architecture of RQUGE metric (upper-side) for the question generation task, which consists of a question answering and a span scorer modules to compute the acceptability of the candidate question. Reference-based metrics are also shown at bottom of the figure, where the score is calculated by comparing the gold and predicted questions.

and predicted sequences in a discrete space by utilising token-level matching functions. Then, recent work e.g. BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) use BERT (Devlin et al., 2019) embeddings to provide a soft token-level matching instead of the hard n-gram overlap. These metrics have been applied to various NLG tasks (Du et al., 2017; Zhou et al., 2017; Xiong et al., 2019; Pan et al., 2020; Lewis et al., 2020; Cheng et al., 2021; Mohammadshahi et al., 2022a,b). Specifically for QG evaluation, Nema and Khapra (2018) propose a scoring function, focusing on the *answerability* of the candidate question, which improves the human correlation when integrated with existing unsupervised metrics.

Regression-based Metrics. These metrics e.g. COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), S3 (Peyrard et al., 2017), and VRM (Hirao et al., 2007) train a regression layer in a supervised manner to mimic the human judgment.

Ranking-based Metrics. The aim of these metrics is to assign a higher score to a better candidate compared to worse predictions. The most popular ones include BEER (Stanojević and Sima'an, 2014) and COMET (Rei et al., 2020).

Generation-based Metrics. The idea is to formulate the evaluation of NLG, as a text generation problem from pre-trained language models. Given the source sequence, the better candidate should be generated with a higher score (probabil-

ity) compared to the worse ones. The most popular ones are BARTScore (Yuan et al., 2021) and PRISM (Thompson and Post, 2020).

Additionally, we include CTC (Deng et al., 2021) and QRelScore (Wang et al., 2022) as reference-free metrics for better comparison. CTC (Deng et al., 2021) proposes an evaluation framework for NLG tasks by providing several reference-free metrics, which are computed by aggregating the alignment scores between the input, context and the predicted sequence. To measure the alignment, CTC (Deng et al., 2021) uses BERT (Devlin et al., 2019) embedding matching, discriminative, and regression models. [3] QRelScore computes the answerability of the candidate question by applying word-level hierarchical matching and sentence-level prompt-based generation. Different from previous work, RQUGE combines question answering and span scoring modules to compute the acceptability of the candidate, which leads to a significantly better correlation with human judgement in multiple datasets with different domains and answer lengths.

3 RQUGE Architecture

RQUGE architecture is illustrated in Figure 2. It consists of two components: question answering and span scorer modules. Given the context, gold answer span, and the candidate question, generated by a question generation model (QG), RQUGE

---

[3]CTC can be considered as an unsupervised metric when BERT embeddings are used to compute the alignment.

computes the acceptance score (κ) of the candidate question as follows:

$$\begin{cases} a_c = \mathrm{QA}(q_c, D) \\ \kappa = \mathrm{S}(q_c, a_c, a_r, D) \end{cases} \quad (1)$$

where the $q_c = \mathrm{QG}(a_r, D)$ is the generated candidate question for the gold answer span a_r and context D. To calculate the score, the question answering model $\mathrm{QA}(.)$ predicts the answer span a_c, given the candidate question q_c and the context D. Finally, the span scorer $\mathrm{S}(.)$ computes the acceptance score κ, conditioned on the candidate question, predicted answer, gold answer, and context. In the following, we will describe each module in detail.

3.1 Question Answering Module

Given the context and the candidate question, the question answering model predicts the answer span. To make our metric general to several domains, we use UnifiedQAv2 (Khashabi et al., 2022) model to generate the answer span. UnifiedQAv2 is a T5-based encoder-decoder model, which is trained on 20 QA datasets, and achieves competitive performance with the state-of-the-art models in several in-domain and out-of-domain datasets.[4] The input to the model is the concatenation of the candidate question and corresponding context.

3.2 Span Scorer Module

Given the predicted answer span a_c of the candidate question q_c, the span scorer calculates the score (ranging from 1 to 5) of the candidate question. Inspired by Chen et al. (2020) and Fabbri et al. (2022), we use an encoder-only BERT-based model to calculate the acceptance score. Specifically, we first encode the input sequence, then pass the vector representation of [CLS] to the regression layer to compute the acceptance score κ. The input to the module is:

[CLS] cand. question [q] gold answer [r] pred answer [c] context

We employ pre-trained RoBERTa model, provided by Fabbri et al. (2022). The model is first pre-trained with a QA-infused pre-training objective,[5]

then fine-tuned on MOCHA human ratings QA dataset (Chen et al., 2020). MOCHA is a dataset of human judgment scores for training and testing reading comprehension metrics, where annotators are asked to score candidate spans, given the context, gold answer, and the corresponding question.

4 Experimental Setup

Datasets. We evaluate metrics on three widely-used QA datasets, including SQuAD(v1) (Rajpurkar et al., 2016), NQ (Kwiatkowski et al., 2019), and MS-MARCO (Bonifacio et al., 2021). NQ is used to demonstrate the benefit of our metric in cases where reference questions are not well-formed and are derived from the Google engine. [6] For MS-MARCO, we use DESCRIPTION type of the dataset to show the effectiveness of our metric on candidate questions with long answer spans (13 tokens on average). Unlike SQuAD and NQ, MS-MARCO is not included in the training data of the question answering module of RQUGE (i.e. UnifiedQAv2 (Khashabi et al., 2022)). We use MS-MARCO to demonstrate that RQUGE can be generalised to out-of-domain datasets. [7]

Question Generators. We fine-tune two commonly used QG models, including GPT2 (Radford et al., 2019), trained with causal language modelling objective, and MixQG (Murakhovs'ka et al., 2022), which is the state-of-the-art question generator and is a T5-based (Raffel et al., 2020) sequence-to-sequence model. We choose GPT2 and MixQG as our question generators as there is a significant gap in their performance, making them suitable for evaluating the metrics. [8]

Baselines. We include BLEU-4 (Papineni et al., 2002), ROUGE-1, ROUGE-L (Lin, 2004), METEOR (Denkowski and Lavie, 2010), Mover-Score (Zhao et al., 2019), and BERTScore (Zhang et al., 2020)[9] as unsupervised metrics, that are commonly used for the question generation task. We additionally use QBLEU, which is specific for

---

information about questions that can be answered by each text span. The pre-trained model is available at `https://github.com/salesforce/QAFactEval`.

[6]We use the subset of NQ with *short-form* answer spans, similarly to the previous work (Murakhovs'ka et al., 2022).

[7]Further details of evaluated datasets are provided in Appendix B.1.

[8]Hyper-parameters for fine-tuning QG models are provided in Appendix B.2.

[9]We use the better-performing BERTScore, initialised with DeBERTa-xlarge model (He et al., 2021).

---

[4]We refer to Khashabi et al. (2022) for further details. A list of evaluated datasets is provided in Appendix A. We use `unifiedqa-v2-t5-large` checkpoint, provided in `https://github.com/allenai/unifiedqa`.

[5]It includes pre-training contextual embeddings with a bi-encoder extractive QA loss, which results in encoding

Metric	Grammaticality			Answerability			Relevance		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
Unsupervised									
BLEU-4	0.133	0.096	0.077	0.273	0.335	0.258	0.213	0.235	0.191
ROUGE-1	0.156	0.096	0.077	0.312	0.274	0.217	0.330	0.322	0.264
ROUGE-L	0.210	0.148	0.120	0.321	0.294	0.233	0.322	0.316	0.259
METEOR	0.143	0.086	0.069	0.334	0.321	0.251	0.317	0.315	0.255
QBLEU	0.160	0.134	0.106	0.227	0.235	0.183	0.240	0.248	0.200
MOVERScore	0.161	0.103	0.082	0.294	0.318	0.248	0.280	0.313	0.254
BERTScore	0.262	0.203	0.160	0.336	0.333	0.260	0.309	0.311	0.253
Regression-based									
BLEURT-20	0.203	0.144	0.113	0.359	0.341	0.268	<u>0.363</u>	<u>0.363</u>	<u>0.295</u>
Ranking-based									
COMET	<u>0.309</u>	<u>0.274</u>	<u>0.215</u>	0.319	0.312	0.243	0.300	0.307	0.248
Generation-based									
BARTScore	0.212	0.145	0.115	0.349	<u>0.345</u>	<u>0.269</u>	0.332	0.323	0.262
Ref-Free									
CTC	0.120	0.131	0.110	0.291	0.243	0.185	0.195	0.179	0.145
QRelScore	0.202	0.102	0.102	<u>0.366</u>	0.285	0.22	0.294	0.212	0.188
RQUGE	**0.380**	**0.278**	**0.220**	**0.604**	**0.436**	**0.344**	**0.551**	**0.403**	**0.325**

Table 1: Correlation of human judgment and automatic evaluation metrics based on Pearson r, Spearman ρ, and Kendall τ correlation coefficients, averaged over subsets of SQuAD, MS-MARCO, and NQ datasets. The best and second best scores are specified with bold and underline markers.

QG evaluation. Furthermore, we utilise BLEURT-20 (Sellam et al., 2020), and COMET (Rei et al., 2020) as regression-based and ranking-based metrics. Finally, we include BARTScore (Yuan et al., 2021), fine-tuned on ParaBank2 (Hu et al., 2019). In all aforementioned metrics, scores are calculated between the candidate and reference questions. As reference-free baselines, we use QRelScore (Wang et al., 2022), and also adopt the factual consistency scorer of CTC (Deng et al., 2021) to calculate the consistency score as:

$$\kappa_{ctc} = \mathrm{mean}(\mathrm{align}([a_r, D] \rightarrow q_c))$$

where a_r, D, and q_c are answer span, context, and candidate question, respectively. The function $\mathrm{align}(.)$ estimates the alignment for tokens of the candidate question with the given context and answer. we use `albert-xlarge-vitaminc-mnli`, which uses a discriminative model to compute the alignment.[10]

Human Annotation. We use 3 volunteer annotators to rate the candidate questions of QG models.[11] All annotators are fluent English speakers.

Inspired by previous work (Rus et al., 2010; Nema and Khapra, 2018), we ask annotators to score each candidate question based on three criteria: *grammaticality*, *answerability*, and *relevance*. Grammaticality measures the syntactic structure of the question. Answerability checks whether the question contains all the important entities, and relevance checks the relatedness of the generated questions with the given answer span. Grammaticality and answerability scores are on a 3-point scale (3 as acceptable, and 1 as rejection), and relevance is on a 2-point scale. We sample 600 questions generated from fine-tuned QG models on SQuAD(v1) (Rajpurkar et al., 2016), NQ (Kwiatkowski et al., 2019), and MS-MARCO (Bonifacio et al., 2021) datasets. We then randomly shuffle and anonymise them for annotators. Further details of the human annotation procedure are provided in Appendix C.[12]

5 Results and Discussion

We evaluate our RQUGE and previous metrics on various datasets and tasks. First, we evaluate the correlation of metrics with human judgment in Sections 5.1 and 5.2. We then demonstrate their robustness on the adversarial subset in Section 5.3.

---

[10]We also used faithfulness aspect of BARTScore (Yuan et al., 2021) as a reference-free metric for measuring the quality of the candidate questions, but preliminary experiments result in a poor correlation with human judgment.

[11]Human evaluation of the quality of the generated questions is not available in previous work.

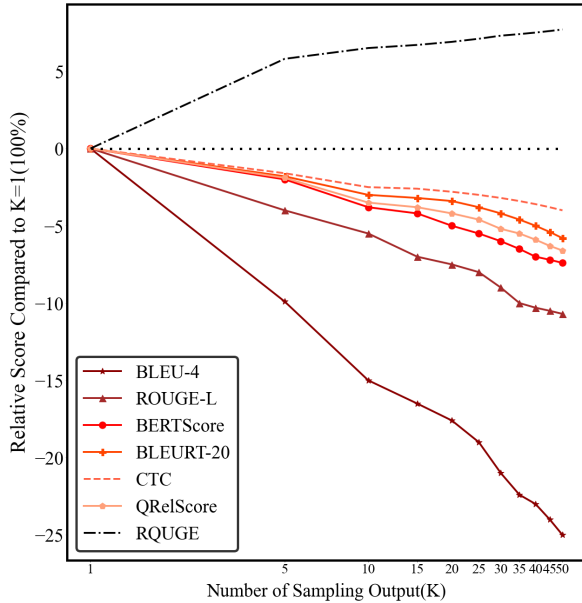[12]The human-evaluated data is available at https://github.com/alirezamshi/RQUGE.

Figure 3: The relative score of automatic metrics compared to $K = 1$ for different values of K, after re-ranking the output of the question generation model with RQUGE metric. RQUGE increases as it is the objective of the re-ranking mechanism.

Finally, Section 5.4 illustrates that fine-tuning QA models on the synthetic data, created by our metric, improves their performance on out-of-domain datasets.

5.1 Correlation with Human Judgment

Annotator Agreement. The pairwise inter-annotator agreements, calculated using Cohen's Kappa are 88.91%, 85.32%, and 83.54%. [13] We use the average score of three annotators for the remaining experiments.

Metric-to-Human Correlation. Table 1 illustrates the correlations of automatic metrics with the human judgment, averaged over all datasets.[14] RQUGE metric has a considerably higher correlation with human judgment on all criteria. For instance, it outperforms the best previous work with +7.1%, +23.8%, +18.8% absolute improvement (based on Pearson (r) score) for *grammaticality*, *answerability*, and *relevance*, respectively. Appendix D illustrates the result of correlation with the human judgment for each dataset, separately. In in-domain evaluation sets, RQUGE results in
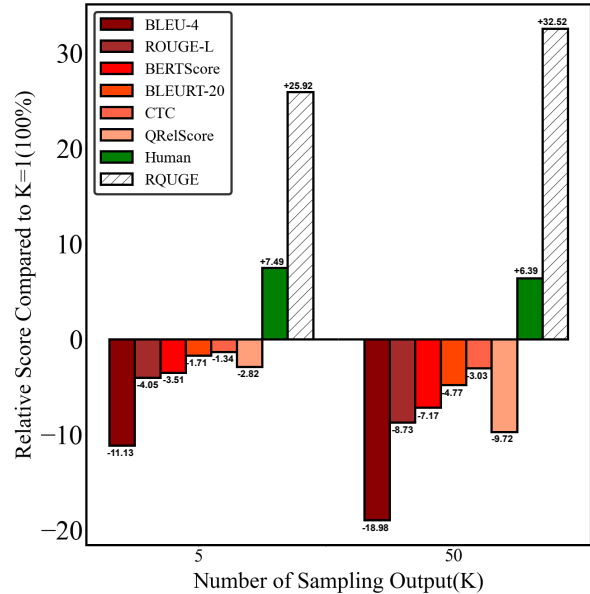

Figure 4: Relative score of automatic metrics compared to $K = 1$ for 250 samples from the evaluation set of SQuAD. For the human score, we use the average score of corresponding samples. It shows that re-ranking with RQUGE gives better output (as both $K = 5, 50$ have better correlation with human), while all QG metrics except RQUGE are diverging after the re-ranking.

+29.7% and +24.8% absolute point improvement for SQuAD and NQ datasets, respectively, based on answerability measurement. For MS-MARCO as the out-of-domain dataset, RQUGE reaches +12.2% absolute improvement for the relevancy criterion, while having competitive results with CTC on answerability measurement. These results show the effectiveness of our metric in different domains, and question structures (well-formedness) and confirm the generalisation of our metric to out-of-domain settings.

5.2 Re-Ranking with RQUGE

To further demonstrate the effectiveness of RQUGE, we use it to re-rank the output predictions of the question generation model to choose the best generated question. Given the context and answer span, QG model [15] generates a bag of candidate questions (here, we apply Nucleus sampling (Holtzman et al., 2020) to increase the diversity)[16], that are sorted based on the perplexity (PPL) of the question generator. At each step, we choose K-

---

[13]Calculated on the summation of grammaticality, answerability, and relevance.

[14]For a fair comparison on *grammaticality*, we just evaluate on SQuAD evaluation set, where reference questions are mostly well-formed.

[15]We use MixQG (Murakhovs'ka et al., 2022) model as the question generation model, as it is the state-of-the-art QG model for the evaluated dataset.

[16]We use temperature sampling of 1, and top summation probability (top_p) of 0.94.

Corruption	Context	Reference	Candidate
Paraphrasing	**Orange County** is a rapidly developing business center that includes Downtown Santa Ana, the South Coast Metro and Newport Center districts; ...	Which county is **developing** its business center?	Which county is **expanding** its business center?
Negation	**Ondemar Dias** is accredited with first discovering the geoglyphs in 1977 and Alceu Ranzi with furthering their discovery after flying over Acre. ...	Who is given credit for discovering geoglyphs along the Amazon River?	Who is **not** given credit for discovering geoglyphs along the Amazon River?
Entity Swap	..., a plague claimed some 1.7 million victims in **Italy**, ... killed about 100,000 in Sweden, and **300,000 in Prussia**. The plague killed ...	How many were killed by plague in **Italy** in the 17th century?	How many were killed by plague in **Prussia** in the 17th century?
Reverse Gender	... For example, **Joseph Haas** was arrested for allegedly sending an email to the Lebanon, New Hampshire city councilors stating, "Wise up or die."	What did Joseph Haas say in **his** email?	What did Joseph Haas say in **hers** email?

Table 2: Samples of the adversarial subset for evaluating the robustness of QG metrics.

first candidates of each sample and re-rank them with RQUGE. Then, other automatic metrics are computed for the best one chosen by RQUGE. Figure 3 demonstrates the relative score gains of QG metrics compared to $K = 1$ (best candidate by PPL) for different values of K. [17] Interestingly, it illustrates that other metrics drop as the number of candidate outputs of each sample (K) is increasing, meaning that the best candidates chosen by our metric are not correlated with other metrics. To confirm these results, annotators are asked to score 250 samples [18] from the evaluation dataset. For each sample, the best candidate questions of $K = 1$ (best predictions of PPL), $K = 5$ (where the RQUGE is starting to become *plateau*), and $K = 50$ (best predictions of RQUGE) are chosen, and annotators are asked to score these three generated questions based on criteria defined in Section 4.[19]

Figure 4 depicts the relative difference of NLG metrics compared to the score of $K = 1$, alongside human scores. We can see from the figure that annotators significantly prefer highest ranking questions of $K = 5$ and $K = 50$ chosen by RQUGE compared to the best candidate questions picked based on PPL, while the average scores of other automatic metrics drop as the number of candidate questions (K) increases. For $K = 5$, the average human score of highest ranking questions is rela-

Context: In 1066, Duke William II of Normandy conquered England killing King Harold II **at the Battle of Hastings**. The invading Normans and their descendants replaced the Anglo-Saxons as …

Reference Where did Harold II die?

Question (*K=1*) When was King Harold II killed?

BLEU-4:14.5 **BERTScore:**0.7 **RQUGE:**1.98 **Human:** 1

Question (*K=50*) Where did Duke William II of Normandy kill King Harold II?

BLEU-4:10.1 **BERTScore:**0.5 **RQUGE:**5.00 **Human:** 2

Figure 5: A sample of re-ranking experiment, that the annotator prefers the best candidate, chosen based on RQUGE ($K = 50$) compared to the question selected based on the perplexity of QG model ($K = 1$).

tively +7.49% better than best candidate questions, chosen by PPL of question generator ($K = 1$). It confirms the effectiveness of RQUGE , when integrated into the decoding step. Additionally, there is not a significant difference (-1.1%) between the average human scores of candidate questions chosen based on RQUGE for $K = 5$ and $K = 50$. This is correlated with Figure 3, as RQUGE also becomes *plateau* from $K = 5$. Figure 5 illustrates an example in which annotators prefer the second candidate question, while BLEU-4 and BERTScore compute higher scores for the first candidate question. [20]

5.3 Robustness Analysis

To further assess the robustness of the QG metrics on adversarial corruptions of reference questions, we evaluate metrics on a subset of positive and negative samples, created from SQuAD (Ra-

---

[17]The relative performances of the remaining metrics for different K are provided in Appendix E.

[18]We filter redundant instances (samples that all of its best candidates are the same) during the sampling process.

[19]Summation of scores for *grammaticality*, *answerability*, and *relevance* is defined as the overall score.

[20]For transparency, we will provide the human evaluation of the re-ranking experiment.

Metric	Total	Neg.	Rev. Gen	Swap Ents
Unsupervised				
BLEU4	0.239	0.241	0.219	0.241
ROUGE-1	0.148	0.126	0.209	0.272
ROUGE-L	0.13	0.11	0.209	0.272
METEOR	0.13	0.09	0.198	0.250
QBLEU	0.220	0.172	0.117	0.558
MOVERScore	0.180	0.169	0.161	0.236
BERTScore	0.408	0.44	0.148	0.285
Regression-based				
BLEURT-20	0.632	0.69	0.24	0.489
Ranking-based				
COMET	0.456	0.523	0.137	0.216
Generation-based				
BARTScore	0.581	0.647	0.205	0.336
Ref-Free				
CTC	0.539	0.576	0.372	0.376
QRelScore	0.546	0.566	**0.420**	0.535
RQUGE	**0.715**	**0.759**	0.371	**0.57**

Table 3: Area under the ROC curve (AUC) of binary classification on the adversarial subset of SQuAD dataset. First column represents the overall performance. Other columns demonstrate AUC metric for different negative sampling methods e.g. negation, reversing the gender, and swapping entities.

jpurkar et al., 2016) evaluation set, as shown in Table 2.[21] The remaining subset contains 2,500 samples equally selected from positive and negative questions. Inspired by Chen et al. (2021) and Honovich et al. (2022), positive questions are paraphrases of references, created by two methods, either translating to a high-resource language, then back-translating to English, or applying a T5 (Raffel et al., 2020) model fine-tuned on Quora paraphrasing task.[22] For negative samples, we use three strategies: negation, reversing the gender, and swapping the entities of the reference question with relevant entities in the corresponding context. Further details of the adversarial evaluation set are provided in Appendix F.

Results and Discussion. We use RQUGE and previous metrics on the adversarial subset to classify the corrupted candidate questions based on their acceptability score. Table 3 illustrates the area of the ROC curve of QG metrics on the adversarial subset. [23] Overall, RQUGE metric significantly

---

outperforms BLEURT-20 (the best previous work) by +13.1% relative improvement. Previous unsupervised metrics drop significantly for all types of negative samples, while BLEURT-20, BARTScore, and reference-free metrics perform better comparatively, especially for *negation*. Our RQUGE metric decreases the error relatively by +22.2% and +7.5% for negation, and entity swapping compared to previous work and has the second-best results on reversing the gender. This confirms the robustness of our metric for different adversarial corruption.

5.4 Domain Adaptation of QA Task

Generated questions using a QG model can be used to improve the performance of a question answering model on out-of-domain datasets. In this section, we show that fine-tuning on the generated synthetic data, re-ranked with RQUGE improves the performance of the question answering model.

Implementation Details. For the out-of-domain dataset, we choose MS-MARCO (Bonifacio et al., 2021) dataset, since the UnifiedQAv2 (Khashabi et al., 2022) (utilised in the calculation of RQUGE) has not used it for training. [24] Given the context, we apply Stanza (Qi et al., 2020a) Named-Entity Recognition (NER) model to extract candidate answer spans. A QG model is then applied to a randomly chosen candidate span, creating a bag of output predictions, using Nucleus sampling (Holtzman et al., 2020). Then, we apply the same re-ranking mechanism, described in Section 5.2 using RQUGE , CTC, QRelScore, and the PPL of the QG. We also use a beam search of size 5 with no re-ranking as a baseline. We use MixQG (Murakhovs'ka et al., 2022) to generate questions. For QA, we first fine-tune T5-small (Raffel et al., 2020) on SQuAD (Rajpurkar et al., 2016) (zero-shot for our setting), then fine-tune it on the generated synthetic data. Further implementation details and hyper-parameters are provided in Appendix G.

Results and Discussion. Figure 6 demonstrates the performance of the QA model on the out-of-domain dataset, fine-tuned for different amounts of synthetic data. Generally, fine-tuned QA model reaches significantly better performance compared to the zero-shot setting. This is important for domains in which we do not have annotated QA data. Furthermore, fine-tuning on the re-ranked data with
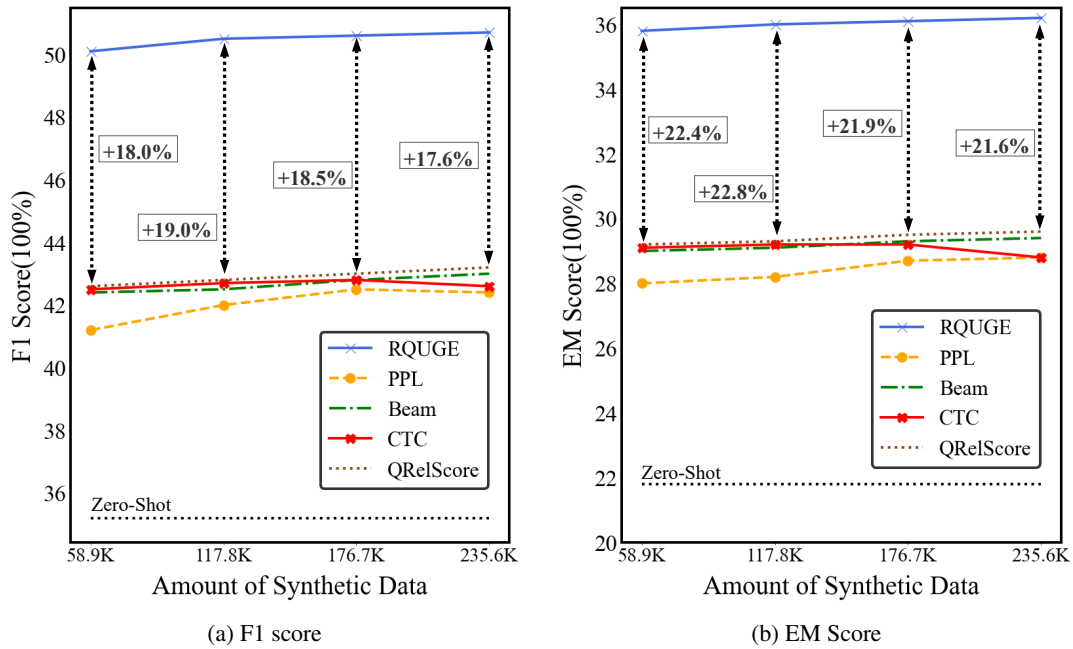

Figure 6: Performance of QA model (based on F1 and Exact Match (EM)) on out-of-domain dataset for different amount of synthetic data. Numbers in the box illustrates the relative improvement of RQUGE compared to the best baseline. 235.6K refers to the setting in which we did not see further improvement using additional synthetic data.

RQUGE consistently improves the performance of the QA model for a different amount of synthetic data, compared to other baselines. Specifically, it significantly outperforms baselines by +18.3% F1, and +22.2% EM, on average. It again shows the effectiveness of our RQUGE by employing it in the domain adaptation of QA models for the out-of-domain dataset.

6 Conclusion

We propose RQUGE , **R**eference-free **QU**estion **G**eneration **E**valuation metric to measure the quality of the generated questions, by better encoding the relevant context and answer without requiring a reference question. It consists of two modules, a question answering model, and a span scorer, which are existing pre-trained models without further fine-tuning. We compare the performance of RQUGE with existing QG metrics on SQuAD, MS-MARCO, and NQ datasets, and show that RQUGE achieves a significantly better correlation with human judgment. Additionally, we integrate RQUGE into the decoding step by using it to re-rank the candidate questions, which leads to a better correlation with human. For robustness, we evaluate QG metrics on adversarial data by corrupting the reference questions and show that RQUGE achieves significantly better performance compared to previous work. Finally, we show that fine-tuning QA models

on the synthetic data, generated with a QG model and re-ranked with RQUGE , improves the performance of QA models on out-of-domain datasets.

Acknowledgement

We thank Parth Pathak, Yatharf Saraf, and Omprakash Sonie for their helpful discussion and support. We are grateful to anonymous reviewers for their fruitful comments and corrections.

Limitations

The main limitation of our work is that we have applied and verified the effectiveness of our metric on the English question answering datasets. Since RQUGE depends on a strong question answering module, one has to find an alternative model to the UnifiedQA (Khashabi et al., 2022) we have used in calculation of RQUGE. Additionally, we did an error analysis on the subset that RQUGE and human evaluation have a significant difference in Appendix H, which shows that mistakes are categorised into syntactic-based and knowledge-based errors. It gives us directions for future improvement of RQUGE metric.

References

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physi-

cal reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *EMNLP*.

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Shaobo Cui, Xintong Bao, Xinxing Zu, Yangyang Guo, Zhongzhou Zhao, Ji Zhang, and Haiqing Chen. 2021. Onestop qamaker: Extract question-answer pairs from text in a one-stop approach.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP*.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352,

Vancouver, Canada. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*. Under review.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding.

Tsutomu Hirao, Manabu Okumura, Norihito Yasuda, and Hideki Isozaki. 2007. Supervised automatic evaluation for summarization with voted regression model. *Information Processing & Management*, 43(6):1521–1535. Text Summarization.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue*

and Conversational Question Answering, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2278–2283, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *ArXiv*, abs/1910.11473.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021. Qace: Asking questions to evaluate an image caption. *arXiv preprint arXiv:2108.12560*.

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.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 742–757. PMLR.

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.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *ArXiv*, abs/1908.05852.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Alireza Mohammadshahi and James Henderson. 2020. Graph-to-graph transformer for transition-based dependency parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3278–3289, Online. Association for Computational Linguistics.

Alireza Mohammadshahi and James Henderson. 2021a. Recursive Non-Autoregressive Graph-to-Graph Transformer for Dependency Parsing with Iterative Refinement. *Transactions of the Association for Computational Linguistics*, 9:120–138.

Alireza Mohammadshahi and James Henderson. 2021b. Syntax-aware graph-to-graph transformer for semantic role labelling.

Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. 2019. Aligning multilingual word embeddings for cross-modal retrieval task. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 11–17, Hong Kong, China. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022a. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022b. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4308–4329, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with

mixed answer types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. 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.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. *arXiv e-prints*, page arXiv:1803.05223.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. 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.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020a. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020b. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

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, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019a. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019b. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 6027–6032.

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.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023a. An empirical comparison of lm-based question and answer generation methods. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023b. A practical toolkit for multilingual question and answer generation, acl 2022, system demonstration. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Toronto, Canada. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022. Qrelscore: Better evaluating generated questions with deeper understanding of context-aware relevance.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2022. QAConv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411, Dublin, Ireland. Association for Computational Linguistics.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A social

media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. 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 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study.

Appendix A   Evaluated Datsets of UnifiedQAv2 Model

UnifiedQAv2 is evaluated on SQuAD (v1) (Rajpurkar et al., 2016), SQuAD (v2) (Rajpurkar et al., 2018), NewsQA (Trischler et al., 2016), Quoref (Dasigi et al., 2019), ROPES (Lin et al., 2019), NarrativeQA (Kočiský et al., 2018), DROP (Dua et al., 2019), NaturalQuestions (Kwiatkowski et al., 2019), MCTest (Richardson et al., 2013), RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), QASC (Khot et al., 2020), PhysicalIQA (Bisk et al., 2019), SocialIQA (Sap et al., 2019), Winogrande (Sakaguchi et al., 2021), BoolQ (Clark et al., 2019), MultiRC (yes/no) (Khashabi et al., 2018), and BoolQ-NP as *in-domain* datasets. Additionally, it is evaluation on AdversarialQA (Bartolo et al., 2020), ReCoRD (Zhang et al., 2018), RACE-C (Liang et al., 2019), HeadQA (Vilares and Gómez-Rodríguez, 2019), MMMLU (Hendrycks et al., 2020), ReClor (Yu et al., 2020), Quail (Rogers et al., 2020), OneStopQA (Cui et al., 2021), MCScript (Ostermann et al., 2018), MCScript 2.0 (Ostermann et al., 2019), CosmosQA (Huang et al., 2019), DREAM (Sun et al., 2019), ProcessBank (Berant et al., 2014), PROST (Aroca-Ouellette et al., 2021), StrategyQA (Geva et al., 2021), PubmedQA (Jin et al., 2019), QAConv (Wu et al., 2022), and TweetQA (Xiong et al., 2019) as *out-of-domain* evaluation sets.

Appendix B   Implementation Details

B.1   Details of Evaluated Datasets

We evaluate QG metrics on three datasets, SQuAD (v1) (Rajpurkar et al., 2016) (under CC BY-SA 4.0 license), Natural Questions (Kwiatkowski et al., 2019) (under Creative Commons Share-Alike 3.0 license), and MS-MARCO (Bonifacio et al., 2021) (fully open-source, no license) datasets. Table 4 illustrates the number of samples in training and evaluation sets.

Dataset	Training Data	Evaluation Data
SQuAD	86,821	5,928
Natural Questions	104,071	12,836
MS-MARCO	502,939	55,578

Table 4: Number of instances for the training and evaluation sets of SQuAD, *short-form* of NQ, and DESCRIPTION types of MS-MARCO datasets.

B.2   Hyper-parameters for Fine-tuning QG Models

All models are trained on NVIDIA A100-SXM4-40GB GPUs. T5 (Raffel et al., 2020) is under Apache License 2.0. GPT2 (Radford et al., 2019) is under modified MIT License. We use AdamW optimiser (Loshchilov and Hutter, 2019), used in several previous works (Mohammadshahi et al., 2019; Devlin et al., 2019; Mohammadshahi and Henderson, 2021b,a, 2020).

Hyper-parameter	Specification
Architecture	T5-base(220M)
No. Encoder Layers	12
No. Decoder Layers	12
No. Epochs	15
Dropout	0.1
Learning rate	3e-5
Batch size	32
No. GPUs	8

(a) MixQG

Hyper-parameter	Specification
Architecture	GPT2(117M)
No. Encoder Layers	12
No. Epochs	12
Dropout	0.1
Learning rate	3e-4
Batch size	32
No. GPUs	8

(b) GPT2

Table 5: Hyper-parameters for fine-tuning QG models on evaluated datasets.

Appendix C   Instruction of Human Evaluation

Annotators are asked to evaluate the quality of a question, given the context and answer span. An input example is shown in Figure 7. They should provide 3 scores for grammaticality, answerability, and relevance. For grammar, the syntactic structure of the sentence is evaluated. They should score 3 for "no grammatical errors", 2 for "not grammatically acceptable but able to get the meaning", and 1 for "unacceptable" questions. For answerability, the score should express the completeness of the candidate question and its consistency with the given answer. So, annotators are required to consider two criteria while scoring; the question should contain question words (e.g. wh-words) and necessary entities, and it should not include the answer. They should score 3, if the question contains all important information, and is consistent with the answer. Score 2 is for the cases, in which some important information is missing in the question or it contains the answer. They should score 1 if all important information is missing in the question and the question is not consistent with the answer.

For relevance, annotators should score the relatedness of the question to the answer, given the context. They should score 2 if the question is answerable by the context and related to the given answer. They should score 1, if the question is out-of-context, or not related to the given answer.

We sample 600 examples (200 for each dataset) from the evaluation sets of SQuAD (v1), NQ, and MS-MARCO. Samples are shuffled and anonymized. All annotators are fluent English speakers.

Context: The Rhine is the longest river in Germany. It is here that the Rhine encounters some more of its main tributaries, such as the Neckar, the Main and, later, the Moselle, which contributes an average discharge of more than **300 m3/s (11,000 cu ft/s)**. Northeastern France drains to the Rhine via the Moselle; smaller rivers drain the Vosges and Jura Mountains uplands. Most of Luxembourg and a very small part of Belgium also drain to the Rhine via the Moselle. As it approaches the Dutch border, the Rhine has an annual mean discharge of 2,290 m3/s (81,000 cu ft/s) and an average width of 400 m (1,300 ft).

Question What is the average discharge of the Moselle?

Grammaticality(1-3):   **Answerability(1-3):**   **Relevance(1-2):**

Figure 7: The input example of the human evaluation.

Appendix D   Correlation with Human Evaluation

Metric	Answerability			Relevance		
	r	ρ	τ	r	ρ	τ
Unsupervised						
BLEU-4	0.256	0.291	0.224	0.213	0.207	0.168
ROUGE-1	0.317	0.292	0.230	0.303	0.281	0.231
ROUGE-L	0.345	0.332	0.263	0.312	0.286	0.235
METEOR	0.337	0.316	0.249	0.308	0.291	0.237
QBLEU	0.296	0.300	0.238	0.272	0.274	0.223
MOVERScore	0.296	0.317	0.248	0.258	0.270	0.218
BERTScore	0.344	0.343	0.266	0.306	0.288	0.233
Regression-based						
BLEURT-20	0.340	0.311	0.242	0.326	0.306	0.247
Ranking-based						
COMET	0.355	0.359	0.279	0.271	0.276	0.222
Generation-based						
BARTScore	<u>0.391</u>	<u>0.383</u>	<u>0.300</u>	<u>0.378</u>	<u>0.334</u>	<u>0.270</u>
Ref-Free						
CTC	0.236	0.130	0.099	0.271	0.189	0.152
QRelScore	0.332	0.276	0.212	0.262	0.213	0.175
RQUGE	**0.688**	**0.388**	**0.303**	**0.588**	**0.404**	**0.327**

Table 6: Correlation of human judgment and evaluation metrics based on Pearson r, Spearman ρ, and Kendall τ correlation coefficients on SQuAD (v1) (Rajpurkar et al., 2016) dataset.

Metric	Answerability			Relevance		
	r	ρ	τ	r	ρ	τ
Unsupervised						
BLEU-4	0.405	0.494	0.398	0.393	0.467	0.380
ROUGE-1	0.533	0.530	0.430	<u>0.517</u>	<u>0.510</u>	<u>0.418</u>
ROUGE-L	0.514	0.523	0.425	0.491	0.491	0.403
METEOR	<u>0.533</u>	<u>0.535</u>	<u>0.430</u>	0.513	0.505	0.411
QBLEU	0.502	0.524	0.417	0.500	0.497	0.405
MOVERScore	0.434	0.482	0.385	0.419	0.480	0.391
BERTScore	0.480	0.490	0.398	0.488	0.495	0.406
Regression-based						
BLEURT-20	0.501	0.506	0.408	0.488	0.509	0.413
Ranking-based						
COMET	0.405	0.393	0.310	0.397	0.403	0.325
Generation-based						
BARTScore	0.421	0.439	0.351	0.419	0.437	0.358
Ref-Free						
CTC	0.270	0.266	0.208	0.270	0.254	0.207
QRelScore	0.415	0.309	0.276	0.394	0.292	0.264
RQUGE	**0.781**	**0.564**	**0.446**	**0.783**	**0.592**	**0.476**

Table 7: Correlation of human judgment and evaluation metrics based on Pearson r, Spearman ρ, and Kendall τ correlation coefficients on Natural Questions (Kwiatkowski et al., 2019) dataset.

Metric	Answerability			Relevance		
	r	ρ	τ	r	ρ	τ
Unsupervised						
BLEU-4	0.222	0.272	0.211	0.096	0.109	0.089
ROUGE-1	0.107	0.086	0.070	0.174	0.199	0.165
ROUGE-L	0.146	0.131	0.106	0.180	0.200	0.165
METEOR	0.168	0.167	0.131	0.167	0.181	0.145
QBLEU	0.138	0.128	0.101	0.134	0.134	0.108
MOVERScore	0.200	0.217	0.168	0.197	0.206	0.167
BERTScore	0.201	0.205	0.159	0.153	0.165	0.134
Regression-based						
BLEURT-20	0.246	0.255	0.202	<u>0.275</u>	<u>0.280</u>	<u>0.229</u>
Ranking-based						
COMET	0.209	0.229	0.181	0.244	0.261	0.213
Generation-based						
BARTScore	0.221	0.236	0.184	0.181	0.207	0.168
Ref-Free						
CTC	**0.401**	**0.376**	<u>0.290</u>	0.154	0.183	0.150
QRelScore	0.351	0.272	0.172	0.226	0.133	0.126
RQUGE	<u>0.400</u>	<u>0.366</u>	**0.293**	**0.397**	**0.356**	**0.288**

Table 8: Correlation of human judgment and evaluation metrics based on Pearson r, Spearman ρ, and Kendall τ correlation coefficients on MS-MARCO (Bonifacio et al., 2021) dataset.
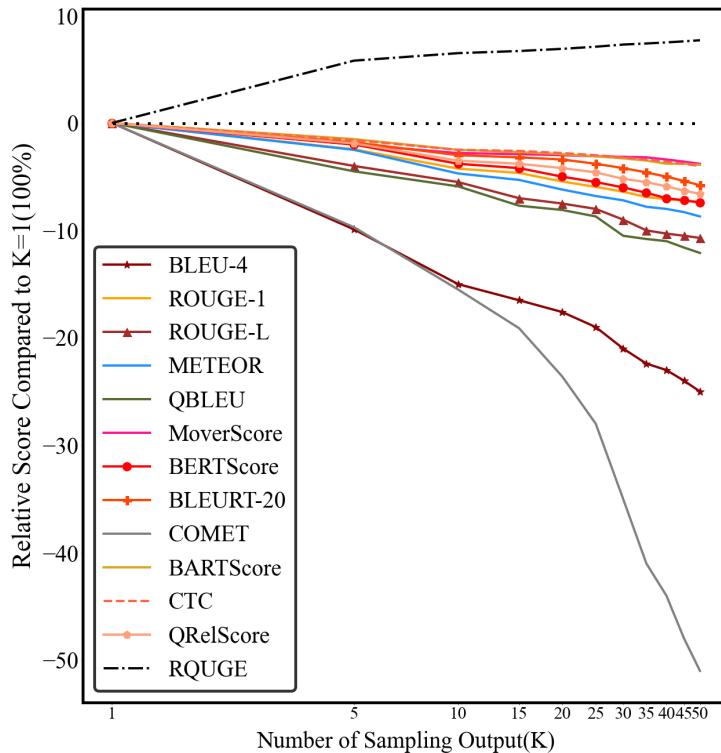
Appendix E   Re-Ranking with RQUGE

Figure 8: The relative score of automatic metrics compared to $K = 1$ for different values of K, after re-ranking the output of question generation model with RQUGE metric. RQUGE increases as it is the objective of the re-ranking mechanism.

Appendix F   Adversarial Evaluation Set

As discussed in Section 5.3, we create positive samples by two mechanisms:

- **Back-Translation.** Translating the reference question to an intermediate language, then translating it back to English. We apply Marian model (Junczys-Dowmunt et al., 2018), and use Chinese and French as intermediate languages, as Marian model has reasonable performance for these language directions.

- **Quora Paraphrasing.** We first train a T5-small (Raffel et al., 2020) model on Quora paraphrasing dataset, [25] and use it for paraphrasing the reference question.

Outputs of both methods are questions that are semantically similar to the reference questions with a few lexical differences.
For the negative samples, as shown in Table 2, we apply the following methods:

- **Negation.** We first scan the reference question to find auxiliary and modal verbs. Then, we randomly either add *not* to the sentence or replace the verb with its antonyms by using WordNet (Miller, 1995) inside the NLTK package (Bird et al., 2009).

- **Reverse Gender.** The reference question is first scanned to find pronouns, and then pronouns are replaced with pronouns with the opposite gender.

---

[25] https://www.kaggle.com/competitions/quora-question-pairs/data

- **Swap Entity.** Stanza (Qi et al., 2020b) named-entity recognition model is applied to the reference question and the context. Then, we randomly select one extracted entity of the reference question and replace it with a random entity of the context with the same entity type.

Appendix G   Implementation Details of Fine-tuning QA models

All models are trained on NVIDIA A100-SXM4-40GB GPUs.

Hyper-parameter	Specification
Architecture	T5-small
No. Encoder Layers	6
No. Decoder Layers	6
No. Training Steps	2K
Dropout	0.1
Learning rate	3e-5
Batch size	32
No. GPUs	8

Table 9: Hyper-parameters for fine-tuning QA models on the synthetic data of MS-MARCO.

Appendix H   Error Analysis

We investigate on cases, in which there is a substantial difference between the human evaluation and RQUGE score. The errors are categorised into syntactic and knowledge-based types, as shown in Table 10. For the syntactic error, RQUGE sometimes computes unacceptable scores for sentences that either miss the question word (e.g. wh-words) or have wrong word order, as QA module of RQUGE focuses more on the semantic aspect of the candidate question to predict the answer span. For the knowledge-based mistakes, RQUGE requires further domain specific and commonsense knowledge to compute the correct score e.g. full moon is instant, not period of time as illustrated in the sample of Table 10. As shown in Table 11, RQUGE computes wrong values for some samples in "reversing gender" and "swapping entities" categories of evaluation set in Section 5.3.

These errors shows the limitations of RQUGE metric, and lead the future work to apply larger and better QA and span scorer modules.

Error Type	Question	Answer & Context	RQUGE	Avg Human
Syntactic				
(1)	cost of wooden shutters	Exterior window shutters cover ... Typical costs: Wooden or vinyl exterior window shutters in stock sizes cost **$20-$200 per pair of panels**.	4.81/5	grammaticality:1.66/3
(2)	SAT solvers routinely handle large instances of what?	... Similarly, algorithms can solve the NP-complete knapsack problem over a wide range of sizes in less than quadratic time and SAT solvers routinely handle large instances of the **NP-complete Boolean satisfiability problem**.	4.76/5	grammaticality:2/3
Knowledge-based	how long is a full moon	**A full lunar cycle lasts almost a month (about 29.5 days)**, and ... However, a full moon, a new moon, and a half moon (first and third quarter) are instants, not periods of time.	4.95/5	answerability:1/3, relevance:1/3

Table 10: Different categories of errors that RQUGE metric computes wrong scores.

Corruption Type	Ref Question	Corrupted Question	Answer & Context	RQUGE
Reversing gender	In what year was the university's 5th president granted his position?	In what year was the university's 5th president granted hers position?	In **1929**, the university's fifth president, Robert Maynard Hutchins, took office; the university underwent many changes during his 24-year tenure...	2.25/5
Swapping entities	The Kuznets curve says with economic development, inequality will decrease after what?	The Piketty curve says with economic development, inequality will decrease after what?	Studies on income inequality and growth have sometimes found evidence confirming the Kuznets curve hypothesis, which states that with economic development, **inequality first increases**, then decreases. Economist Thomas Piketty challenges this notion...	4.65/5

Table 11: Some samples from the adversarial subset of Section 5.3, that RQUGE metric is not sensitive to the corruption.

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section and Appendix H.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. We provide an evaluation metric for QG task. limitations are provided in Conlusion and Limitations sections.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We include our contributions in both abstract and introduction sections. Specifically, we provide them at the end of introduction section.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

For the model, we use T5 and GPT2 in our paper (section 3 and section 5). For the metric, we use different automatic evaluation metric (sections 2,4,5). For the dataset, we use SQuAD, NQ, and MS-MARCO (sections 4 and 5). We also create the human annotation for the evaluation (sections 4 and 5, Appendix C).

☑ B1. Did you cite the creators of artifacts you used?
Sections 4, 2, 3.1 and 3.2.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendices B.1 and B.2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Since our annotations are created from SQuAD, NQ, and MS-MARCO datasets, we use the same license for the distribution of our human annotations (Appendix C). For pre-trained models, we use publicly available models (Appendix B.2).

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We use SQuAD, NQ, and MS-MARCO datasets, which organizers already checked these concerns before making them available.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendices B.1, C, and F.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendices B.1, C, and F.

C ☑ **Did you run computational experiments?**

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendices B.2 and G.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendices B.2 and G.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. For the evaluation of metrics, we use pearson, spearman, and kendall metrics to find the correlation with the human judgment. For QA experiment, we run F1 and EM metrics for the evaluation. In both cases, there were a significant different between our model and previous works. For both QG and QA models, results are single-run, as mentioned in the paper.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and Appendix F.

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

We collect the data for the evaluation of question generation metrics: Section 4 and Appendix C.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
The annotators were volunteer fluent English speaking students (mentioned in Appendix C), and we created our internal website to get the annotations.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. We inform them that the data will be used for the evaluation of question generation task (Appendix C).

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We use the same protocol as previous work (which was approved) in the question generation task.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.