QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation

Anonymous ACL submission

Abstract

Automatically generated questions often suffer from problems such as unclear expression or factual inaccuracies, requiring a reliable and comprehensive evaluation of their quality. Human evaluation is widely used in the field of question generation (QG) and serves as the gold standard for automatic metrics. However, there is a lack of unified human evaluation criteria, which hampers consistent and reliable evaluations of both QG models and automatic metrics. To address this, we propose **QGEval**, a multi-dimensional **Eval**uation benchmark for **Q**uestion **G**eneration, which evaluates both generated questions and existing automatic metrics across 7 dimensions: fluency, clarity, conciseness, relevance, consistency, answerability, and answer consistency. We demonstrate the appropriateness of these dimensions by examining their correlations and distinctions. Through consistent evaluations of QG models and automatic metrics with QGEval, we find that 1) most QG models perform unsatisfactorily in terms of answerability and answer consistency, and 2) existing metrics fail to align well with human judgments when evaluating generated questions across the 7 dimensions. We expect this work to foster the development of both QG technologies and their evaluation.

1 Introduction

Question Generation (QG) is a typical Natural Language Generation (NLG) task that aims to generate natural language questions based on an input context and optionally an answer. QG has broad applications such as question answering (QA) (Lyu et al., 2021), conversational systems (Zeng et al., 2023), and knowledge assessment (Ghanem et al., 2022). However, it has been demonstrated that questions generated by QG models suffer from problems like ambiguities and hallucinations (Laban et al., 2022), which emphasizes the critical importance of reliable evaluations.

Passage: ..... *The publication of a Taoist text inscribed with the name of Töregene Khatun, Ögedei's wife, is one of the first printed works sponsored by the Mongols.....* **Answer:** *Töregene Khatun* **Reference:** *Who was Ögedei's wife?*
Q1: *Who was the name of Ögedei's wife?* **Scores:** Flu. - 2.6667; Clar. - 3; Conc. - 3; Rel. - 3; Cons. - 3; Ans. - 3; AnsC. - 3
Q2: *Who was the Mongol ruler whose name was inscribed on one of the first printed works sponsored by the Mongols?* **Scores**: Flu. - 3; Clar. - 3; Conc. - 3; Rel. - 3; Cons. - 1; Ans. - 1.3333; AnsC. - 1.3333
Q3: *Who was a Taoist text inscribed with the name of gedei's wife?* **Scores:** Flu. - 2.3333; Clar. - 1.3333; Conc. - 3; Rel. - 3; Cons. - 1; Ans. - 1; AnsC. - 1
.....

Table 1: An example of QGEval, including a passage, an answer, a reference question, and 15 generated questions (only 3 are shown for brevity). The score ranges from 1 to 3 (higher better). Errors within questions are highlighted with underlines. Abbreviations are as follows. Flu.:Fluency; Clar.:Clarity; Conc.:Conciseness; Rel.:Relevance; Cons.:Consistency; Ans.:Answerability; AnsC.:Answer Consistency.

Human evaluation is widely acknowledged as the gold standard for evaluating QG (Wang et al., 2022), with most automatic metrics striving to align their results with human evaluation results (Amidei et al., 2018; Sai et al., 2022). However, the criteria of human evaluations are varied in existing research, leading to inconsistent and unreliable evaluations of QG models (Ji et al., 2022) and automatic metrics (Amidei et al., 2018; Mulla and Gharpure, 2023). This inconsistency highlights the urgent need to establish a unified human evaluation benchmark to ensure reliable evaluations.

Despite the importance of such benchmarks, few have been published, and the existing ones usually

have the following limitations: 1) focusing only on specific dimensions like answerability; 2) involving a small amount of data (e.g., <1k samples); 3) employing a limited variety of models to generate questions, resulting in a lack of diversity in the data. For instance, Nema and Khapra (2018) generated monotonous questions using rule-based methods for evaluation and focused primarily on the answerability of questions, neglecting other dimensions. Gollapalli and Ng (2022) evaluated generated questions from four dimensions but only included 500 questions generated by three models. Laban et al. (2022) utilized seven QG models to generate 1k+ questions to be evaluated, however, they merely assessed whether the generated questions could be accepted as reading comprehension quiz questions, rather than scoring them on multiple dimensions.

To address the above issues, we propose QGEval, a multi-dimensional evaluation benchmark, which evaluates questions across 7 dimensions and contains 3k questions generated by 15 QG models (including LLMs) based on 200 passages and answers. Specifically, through preliminary error analysis of the generated questions (described in section 2.2), we identified seven evaluation dimensions and categorized them into two aspects: **1) Linguistic dimensions,** including fluency, clarity, and conciseness, which are basic requirements that a natural language text should meet; and **2) Task-oriented dimensions,** including relevance, consistency, answerability, and answer consistency, which involve requirements specific to QG tasks.

As illustrated in Table 1, both linguistic and task-oriented dimensions are essential for a comprehensive evaluation of generated questions. In particular, although Q1 receives high scores in all task-oriented dimensions, it has a lower fluency score in linguistic dimensions due to the incorrect use of the interrogative word. On the contrary, Q2 performs well across all linguistic dimensions but scores poorly on most task-oriented dimensions because of its inconsistencies with the passage. These examples demonstrate the necessity of the two categories of evaluation dimensions.

Using QGEval to evaluate the performance of 15 different QG models, we find that these models perform relatively poorly in terms of answerability and answer consistency compared to other dimensions. We also evaluate and compare the performance of 15 existing automatic metrics, observing that there is still a gap between these metrics and human evaluations.

To summarize, our main contribution is fourfold:

- We introduce a multi-dimensional evaluation benchmark for QG named QGEval, which assesses the quality of questions across 7 dimensions and contains 3k questions generated by 15 QG models.
- We conduct a detailed analysis of the generated questions and compare the generation performance of various QG models across the seven dimensions, discovering that most models underperform in answerability and answer consistency.
- We evaluate and compare the performance of 15 automatic metrics across the seven dimensions, highlighting the discrepancies between automatic metrics and human evaluation.
- We have made the QGEval dataset, along with the codes for the automatic metrics we utilized, publicly accessible for further research.[1]

2 The QGEval Dataset

In this section, we describe how we construct the QGEval dataset, the overall pipeline includes two stages: question generation and human evaluation, as shown in Figure 1.

2.1 Question Generation

In the first stage, our goal is to generate questions for evaluation. We use SQuAD (Rajpurkar et al., 2016) and HotpotQA (Yang et al., 2018) as the base datasets, which are two widely used datasets in the field of QA and QG. We divide the SQuAD dataset into train/dev/test splits following (Zhou et al., 2018). As for the HotpotQA dataset, we utilize its official train split and designate the first 3700 samples from the official dev set as our dev split and the rest as the test split. The train and dev splits are used to train QG models and the test split is then utilized to generate questions. We randomly select 100 samples from the test split of each dataset and utilize the passage and answer pairs provided by these samples to generate questions. The process results in the dataset to be evaluated, which comprises 3000 questions generated

[1]Our data and code are publicly available at https://anonymous.4open.science/r/QGEval-anonymous-2B32

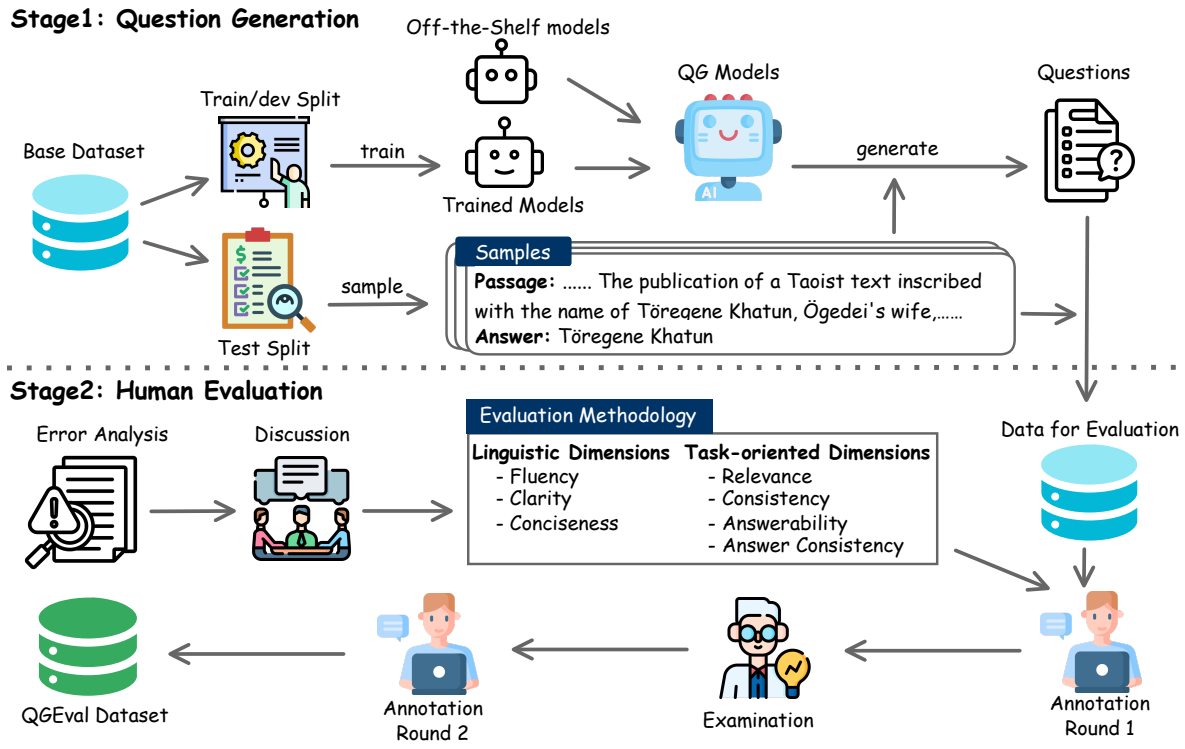


Figure 1: Pipeline of dataset construction. Stage 1: Generate questions to be evaluated. Stage 2: Conduct two rounds of annotation to form the QGEval dataset.

by multiple QG models based on 200 passages and answers. QG models contain both Off-the-Shelf models (public ones already trained on the QG task) and models trained by ourselves, the implementation details of QG models are in Appendix B.

To capture a wide diversity of model outputs and facilitate comparisons between different models and settings, our selection of QG models covers a variety of model sizes, types, and settings. Specifically, we utilize 14 QG models based on different language models and under various settings. The language models cover a broad range of sizes and encompass four different series of models: BART (Lewis et al., 2020), T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2024), and GPT (OpenAI[2]). Settings include fine-tuning, low-rank adaptation (LoRA) (Hu et al., 2022), few-shot, and zero-shot. We customize the settings for models of different sizes, ensuring that each model is equipped with settings suitable for its characteristics. We also regard the references as outputs from one model for subsequent annotation, along with those from the other 14 models. Table 2 shows all language model variants, the number of models' parameters, and the settings we employed for each model.

[2]https://platform.openai.com/docs/models

Model	Param.	Settings
BART-base	140M	fine-tuning
BART-large	400M	fine-tuning
T5-base	250M	fine-tuning
T5-large	780M	fine-tuning
Flan-T5-base	250M	fine-tuning
Flan-T5-large	780M	fine-tuning
Flan-T5-XL	3B	LoRA;few-shot(8)
Flan-T5-XXL	11B	LoRA;few-shot(8)
GPT-3.5-turbo	—	few-shot(8);zero-shot
GPT-4	—	few-shot(8);zero-shot

Table 2: Language models and settings used for question generation. GPT-4 refers to GPT-4-1106-preview. Since the parameter sizes of GPT-3.5-turbo and GPT-4-1160-preview have not been officially announced, we do not list them here.

2.2 Human Evaluation

In the second stage, our objective is to obtain human ratings for each generated question. The evaluation methodology and the process of human annotation will be described in detail.

Evaluation Methodology To figure out which dimensions we should evaluate questions on, we conducted a pilot experiment to analyze the errors presented in the generated questions (see details in

Appendix E.1). We observed that QG models may generate questions that are incorrectly formed (e.g., not a question) and phrased, ambiguous, or verbose, making it difficult to understand their intent. QG models may also generate questions that are irrelevant to the context, inconsistent with the provided information, unanswerable, or mismatched with the given answer, failing to meet the requirements of the QG task. From these observations, we conclude that the errors can be categorized into two types: linguistic and task-oriented. After a thorough discussion with two experts in the field of education, we determined that the quality of questions should be evaluated on the following seven dimensions, including both linguistic and task-oriented aspects.

Linguistic dimensions serve as the foundational evaluation dimensions in most NLG tasks including QG. Specifically, we focus on the following three linguistic dimensions in our evaluation, requiring the generated questions to be well-formed, and expressed clearly and concisely.

- **Fluency (Flu.):** Whether the question is well-formed, grammatically correct, coherent, and fluent enough to be understood (Oh et al., 2023).
- **Clarity (Clar.):** Whether the question is expressed clearly and unambiguously, avoiding excessive generality and ambiguity, the same as the definition in (Ousidhoum et al., 2022).
- **Conciseness (Conc.):** Whether the question is concise and not abnormally verbose with redundant modifiers, as defined in (Cheng et al., 2021).

Task-oriented dimensions refer to those aspects associated with the QG task, measuring the correlation between the generated questions and passages, as well as the connection between questions and the provided answers. The task-oriented dimensions we considered are outlined below, requiring the generated questions to be contextually relevant and consistent, answerable based on the passage, and match the provided answers.

- **Relevance (Rel.):** Whether the question is relevant to the given passage and asks for key information from the passage. It is also a commonly used dimension in both QG and other text generation tasks (Oh et al., 2023; Sai et al., 2022).
- **Consistency (Cons.):** Whether the information presented in the question is consistent with the passage and without any contradictions or hallucinations, similar to the definition in other text generation tasks (Honovich et al., 2022).
- **Answerability (Ans.):** Whether the question can be distinctly answered based on the passage, a widely used and distinctive dimension in QG (Ghanem et al., 2022).
- **Answer Consistency (AnsC.):** Whether the question can be answered using the provided answer, as "Answer Matching" defined in (Cheng et al., 2021).

The scoring scale for each dimension is 1 to 3, with higher being better (details of the scoring criteria are presented in the Appendix A.1)

Annotation Process Due to the subjective nature of annotation, a crowdsourcing approach was adopted. Three postgraduate students specializing in computer science volunteered as annotators to score the generated questions according to the scoring criteria for each dimension on our annotation platform (see the interface in Appendix A.2). Two rounds of annotation were performed to confirm judgments and ensure a higher quality of annotations.

In the first round, questions generated based on SQuAD and HotpotQA were presented and scored separately. For the same passage, all 15 questions generated by different models were presented simultaneously, and annotators scored these questions sequentially. Annotating in this way increases efficiency and helps annotators validate their judgments (e.g., similar questions should receive similar scores). During annotation, the generative models were kept unaware. Annotation results from the first round were examined. In the second round, the annotators were required to review samples that may have been incorrectly scored. For each dimension, the annotators: 1) checked annotations when the same questions received different scores on the same dimension; 2) reviewed samples where their annotations differed from the other annotators by 2 points, while the annotations of the other two annotators were the same; 3) discussed with each other when the annotation scores in the first round were 1, 2, 3.

To assess the agreement between annotators, Krippendorff's alpha coefficient, a statistical mea-

Rounds	Flu.	Clar.	Conc.	Rel.	Cons.	Ans.	AnsC.
Round1	0.226	0.375	0.515	0.233	0.181	0.354	0.559
Round2	0.427	0.576	0.755	0.437	0.445	0.661	0.800

Table 3: Krippendorff's alpha coefficient of inter-annotator scores in the first and second round annotations. A higher score means higher agreement among the annotators.

sure of inter-rater reliability, was calculated for each dimension and shown in Table 3. In the first round, the coefficients ranged from 0.181 to 0.559 and improved to a range of 0.427 to 0.800 in the second round. Furthermore, to verify the quality of annotations, 100 samples were randomly selected and reviewed by the two experts. The results showed that the accuracy of annotations for each dimension was over 96%.

3 Experiment and Evaluation

In this section, we conduct a series of analytical experiments and evaluations on QG models and automatic metrics with QGEval. We aim to address the following three research questions.

3.1 Are the seven dimensions appropriate for the evaluation of QG?

To figure out whether the dimensions are an appropriate set, we examine the correlations and distinctions among them by calculating Pearson correlations and conducting the Nemenyi test on the annotation scores for these dimensions. Pearson correlation measures the linear correlation between two sets of data, with higher absolute values indicating stronger correlations. The Nemenyi test determines whether there are significant differences between groups, with lower p-values indicating greater significance. Intuitively, the seven dimensions might correlate with each other but should also maintain differences from one another. As shown in Figure 2, the Pearson correlations between the seven dimensions are within a reasonable range (0.04 to 0.67), and most p-values in the Nemenyi test are below 0.05. This indicates that **these dimensions are interrelated but still exhibit distinct characteristics**, consistent with our intuition.

For correlations between dimensions, we further observe: 1) The correlation coefficients among the linguistic dimensions (fluency, clarity, and conciseness) are relatively high. 2) Linguistic dimensions can influence task-oriented dimensions. For instance, clarity and consistency show high correlations with answerability. Unclear expression (low clarity) and contradictions between the question and passage (low consistency) may lead to a low score of answerability. 3) As expected, answer consistency is highly relevant to answerability, and from experience, unanswerable questions tend to have low answer consistency scores.

3.2 How do the QG models perform across the seven dimensions?

By asking this question, we aim to explore which dimensions QG models perform well or poorly on and to compare the generation performance of different QG models. Table 4 shows the averaged annotation scores along seven evaluation dimensions of all QG models. Generally speaking, most QG models are capable of generating questions that are both fluent and relevant to the provided passage, i.e., **performing well on both fluency and relevance dimensions.** However, they often **encounter challenges in generating questions that are answerable and align well with the given answers.** Inspired by this finding, we advocate that future question generation work should focus more on improving the answerability and answer consistency of generated questions.

We further compare these models via different model sizes and settings, our findings are: 1) The best three QG models ranked by the average scores of all dimensions are GPT-4-fewshot, GPT-4-zeroshot, and reference, indicating that the quality of questions generated by GPT-4 is comparable to that of humans. 2) Under the same setting, as the model size increases, the generated questions exhibit improved clarity in expression, higher consistency with the provided passages, and increased alignment with the provided answers. 3) Maintaining the same model, the zero-shot approach performs less effectively than the few-shot approach, and the few-shot approach is inferior to the supervised (LoRA) approach, especially on the consistency, answerability, and answer consistency dimensions. 4) Models under zero-shot and few-shot settings often fail to generate questions that match the given answers, except for GPT-4, which could be due to the models' insufficient ability to follow

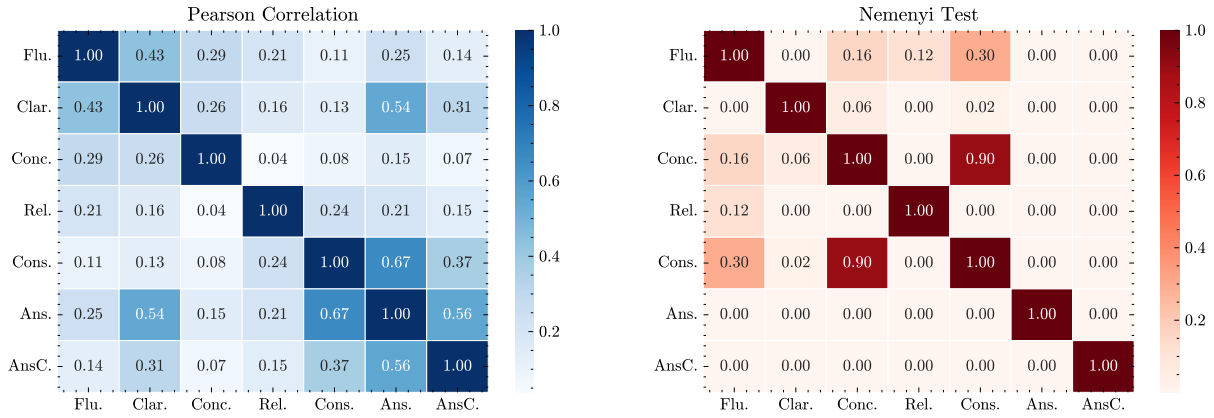


Figure 2: Pearson correlations and p-values of Nemenyi test for seven dimensions.

Models	Flu.	Clar.	Conc.	Rel.	Cons.	Ans.	AnsC.	Avg.
Reference	2.968	2.930	**2.998**	2.993	2.923	2.832	**2.768**	**2.916**
BART-base-finetune	2.958	2.882	2.898	2.995	2.920	2.732	2.588	2.853
BART-large-finetune	2.932	2.915	2.828	2.995	2.935	2.825	**2.737**	2.881
T5-base-finetune	2.972	2.923	2.922	**3.000**	2.917	2.788	2.652	2.882
T5-large-finetune	2.978	2.930	2.907	2.995	2.933	2.795	2.720	2.894
Flan-T5-base-finetune	2.963	2.888	2.938	**2.998**	2.925	2.775	2.665	2.879
Flan-T5-large-finetune	2.982	2.902	2.895	2.995	**2.950**	2.818	2.727	2.895
Flan-T5-XL-LoRA	2.913	2.843	2.880	2.997	2.928	2.772	2.667	2.857
Flan-T5-XXL-LoRA	2.938	2.848	2.907	**3.000**	2.943	2.757	2.678	2.867
Flan-T5-XL-fewshot	2.975	2.820	**2.985**	2.955	2.908	2.652	2.193	2.784
Flan-T5-XXL-fewshot	**2.987**	2.882	**2.990**	2.988	2.920	2.687	2.432	2.841
GPT-3.5-turbo-fewshot	2.972	2.927	2.858	2.995	**2.955**	**2.850**	2.335	2.842
GPT-4-fewshot	**2.988**	**2.987**	2.897	2.992	**2.947**	**2.922**	**2.772**	**2.929**
GPT-3.5-turbo-zeroshot	**2.995**	**2.977**	2.913	2.992	2.917	2.823	2.157	2.825
GPT-4-zeroshot	2.983	**2.990**	2.943	2.970	2.932	**2.883**	2.723	**2.918**
Avg.	2.967	2.910	2.917	2.991	2.930	2.794	2.588	

Table 4: Annotation scores of questions along seven dimensions, averaged over three annotators. The three highest and lowest scores of each dimension are bolded and underlined, respectively. GPT-4 refers to GPT-4-1106-preview. Avg. refers to the average score.

detailed instructions.

3.3 Can existing automatic metrics accurately evaluate generated questions?

In this section, we use QGEval to evaluate and compare the performance of existing automatic metrics to find out whether these metrics are able to accurately evaluate the quality of generated questions across the seven dimensions.

Automatic Metric Our selection of automatic metrics varies from methods based on lexical overlap to those based on large language models, including both reference-based and reference-free approaches. Reference-based metrics evaluate questions by computing the similarity between them and the references, which include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019), BLEURT (Sellam et al., 2020), Q-Metric (Nema and Khapra, 2018), and QSTS (Gollapalli and Ng, 2022). Reference-free metrics utilize the comprehension and generation capabilities of language models to evaluate questions without references, including BARTScore (Yuan et al., 2021), GPTScore (Fu et al., 2023), UniEval (Zhong et al., 2022), QRelScore (Wang et al., 2022), and RQUGE (Mohammadshahi et al., 2023). When using the ref-hypo scoring type (generate candidate text based on the reference), BARTScore and GPTScore are considered reference-based metrics. Among all these metrics, UniEval and GPTScore are designed for multi-dimensional evaluation, offering a score for each dimension (7 scores for 7 dimensions), while the other metrics provide only a single overall score. Detailed descriptions of these metrics are presented in Appendix D.

Metric Evaluation We evaluate the agreement between the automatic metrics and human annotation scores by calculating the Pearson correlation over each dimension, results are shown in Table 5, with the three highest and lowest absolute coefficients bolded and underlined respectively.

Correlation results show several trends. 1) **Most metrics have relatively low correlations with annotation scores along seven dimensions,** ranging from -0.4 to 0.4, especially on fluency, clarity, relevance, and consistency. We observed that most questions received high annotation scores across these four dimensions, while the scores assigned by automatic metrics varied significantly, resulting in poor alignment with human scores. 2) In general, **reference-free metrics tend to outperform reference-based metrics,** exhibiting higher correlation coefficients with human evaluation. 3) **Metrics that conduct multi-dimensional evaluations tend to perform better across a wider range of dimensions** compared to those that provide only a single composite score (BLEU, ROUGE, BARTScore, etc.). UniEval, for example, achieves the three highest coefficients across six dimensions. 4) **Metrics designed for specific dimensions are better than other metrics on those specific dimensions.** RQUGE, leveraging question-answering results for evaluation, attains higher correlations on its target dimensions: answerability and answer consistency. The observations in 3) and 4) imply that metrics with a single composite score are not suitable for the comprehensive evaluation of generated questions. Instead, **designing multi-dimensional metrics or metrics focused on specific dimensions may yield better results.**

Our further exploration of the score distribution of automatic metrics (in Appendix E.3) and the application of these metrics to rank different QG models (in Appendix E.4) indicates that existing automatic metrics still struggle to effectively distinguish questions of varying quality.

LLM as Evaluator Recent work has leveraged LLMs for NLG evaluation and found that LLM-based metrics are superior to former metrics (Kocmi and Federmann, 2023). To assess the effectiveness of employing LLMs for question generation evaluation, we use the GPT-3.5-turbo and GPT-4-1106-preview as evaluators and implement evaluations using both direct prompts and G-EVAL (Liu et al., 2023), an evaluation method employing Chain-of-Thought (COT). Due to budget constraints, we conducted tests on 450 questions (30 passages) solely focusing on the answerability dimension for analysis. The Pearson correlations between annotation scores and metrics are shown in Table 6.

We compare the performance of LLM-based metrics with RQUGE, UniEval, and GPTScore-src (the top three metrics on answerability in Table 5) here. The results show that metrics based on GPT-4 achieve the highest correlations with human scores, which demonstrates the potential of using LLMs for QG evaluation. The comparisons between methods using direct prompts and G-EVAL also verify the effectiveness of COT. Although LLM-based metrics outperform other evaluation methods, they still fail to align closely with human evaluation (Pearson correlations are below 0.4). Further exploration is needed in future work.

4 Related Work

Automatic Metrics Automatic evaluation of QG is still dominated by reference-based metrics such as BLEU (Papineni et al., 2002) and Q-Metric (Nema and Khapra, 2018), which compute the similarity between generated questions and references. As QG is a one-to-many generation task, this type of metric can not evaluate questions that are different from the references (Mohammadshahi et al., 2023). Reference-free metrics like BARTScore (Yuan et al., 2021) and QRelScore (Wang et al., 2022) overcome this limitation, but they often assign a single overall score as the evaluation result, which is less interpretable and not comprehensive. UniEval (Zhong et al., 2022) and GPTScore (Fu et al., 2023) are designed to evaluate generated texts from multiple interpretable dimensions, but they are not specifically designed for the QG task, and thus their performance in evaluating QG is limited.

Human Evaluation in QG Since existing automatic metrics are not effective enough to measure the quality of generated questions, human evaluation is frequently used in the field of QG (Mulla and Gharpure, 2023). However, the human evaluation criteria provided by existing works are disparate, leading to inconsistent evaluation of generated questions. Ghanem et al. (2022) utilized answerability, fluency, and grammaticality to assess question quality, while Ushio et al. (2022) employed grammatically, understandability, and answerability for evaluation. Gou et al. (2023) fo-

Metrics	Flu.	Clar.	Conc.	Rel.	Cons.	Ans.	AnsC.
Reference-based Metrics							
BLEU-4	0.028	0.049	0.138	0.041	0.032	0.080	0.162
ROUGE-L	0.080	0.086	0.234	0.085	0.079	0.127	0.233
METEOR	0.020	0.088	0.106	0.079	0.059	0.131	0.253
BERTScore	0.140	**0.123**	**0.313**	**0.113**	0.091	0.131	0.231
MoverScore	0.070	0.075	0.209	0.071	0.058	0.101	0.188
BLEURT	0.078	**0.105**	0.179	0.104	0.098	0.144	**0.271**
BARTScore-ref	0.087	0.079	0.235	0.109	0.078	0.092	0.190
GPTScore-ref	0.069	0.086	0.182	0.006	0.054	0.106	0.187
Q-BLEU4	0.072	0.082	0.216	0.058	0.075	0.113	0.198
QSTS	0.016	0.104	0.015	0.077	0.043	0.130	0.250
Reference-free Metrics							
BARTScore-src	**-0.148**	-0.035	**-0.511**	0.053	-0.001	0.018	-0.015
GPTScore-src	0.134	0.104	-0.052	**0.416**	**0.197**	**0.148**	0.236
UniEval	**0.370**	**0.219**	0.259	**0.153**	**0.156**	**0.207**	**0.356**
QRelScore	**-0.213**	-0.096	**-0.553**	0.032	0.002	-0.026	-0.025
RQUGE	0.045	0.092	0.126	0.070	**0.200**	**0.211**	**0.561**

Table 5: Pearson correlation between automatic metrics and human scores along seven dimensions. The three highest and lowest absolute coefficients of each dimension are bolded and underlined, respectively. BLEU-4: 4-gram variant of BLEU; ROUGE-L: the longest common subsequence (LCS) variant of ROUGE; *-ref: ref-hypo scoring type, *-src: src-hypo scoring type.

Metrics	Pearson	Metrics	Pearson
GPTScore	0.187	GPT-3.5	0.195
UniEval	0.215	G-EVAL$_{\text{GPT-3.5}}$	0.228
RQUGE	0.250	**GPT-4**	**0.296**
		G-EVAL$_{\text{GPT-4}}$	**0.356**

Table 6: Pearson correlation between annotation scores and metrics on answerability. GPT-3.5 and GPT-4 refer to the methods using direct prompts. GPTScore refers to GPTScore-src.

cused on consistency and diversity of generated questions. Disparate human evaluation criteria also result in inconsistent evaluation and comparison of automatic metrics. Nema and Khapra (2018) proposed Q-metric and computed the correlations between existing automatic metrics and human judgments on answerability. Wang et al. (2022) proposed QRelScore and compared it with other metrics based on their human evaluation results on three dimensions: grammaticality, relevance, and answerability. Mohammadshahi et al. (2023) evaluated the performance of automatic metrics on their newly annotated data as the human evaluation of generated questions is not available in previous work. Thus, it's urgent to develop unified and reliable human evaluation benchmarks to ensure consistent and accurate assessments of generated questions and automatic metrics.

5 Conclusion

In this work, we introduced a comprehensive, multi-dimensional evaluation benchmark, QGEval, to facilitate the evaluation of generated questions from various models and existing automatic metrics across 7 dimensions: fluency, clarity, conciseness, relevance, consistency, answerability, and answer consistency. It contains 3k questions generated from 15 different QG models. Through analysis of QGEval, we found that most models performed unsatisfactorily on answerability and answer consistency. This highlights the importance of focusing on the two dimensions in future QG model designs. Additionally, our evaluation of 15 existing automatic metrics revealed that these metrics still exhibit relatively low correlation coefficients with human annotation scores, emphasizing the need to explore advanced metrics that align better with human evaluation. We hope that this work will serve as a valuable resource for future research on question generation evaluation and models.

6 Limitations

Our work proposes QGEval, a multi-dimensional evaluation benchmark for QG, to evaluate and compare the performance of different QG models and existing automatic metrics. Although it provides a comprehensive evaluation of generated questions, it focuses on the scenario of generating questions based on a passage and an optional answer and

is not applicable to other scenarios such as visual question generation (Vedd et al., 2022) and conversational question generation (Zeng et al., 2023). Moreover, additional dimensions may be introduced to meet some specific requirements. For example, complexity is considered when the generated questions are required to involve multi-hop reasoning (Fei et al., 2022). In this work, we consider more general requirements under the scenario we focus on.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*.

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.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. 2022. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906, Dublin, Ireland. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146.

Sujatha Das Gollapalli and See-Kiong Ng. 2022. QSTS: A question-sensitive text similarity measure for question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3835–3846, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qi Gou, Zehua Xia, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Nguyen Cam-Tu. 2023. Diversify question generation with retrieval-augmented style transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1690, Singapore. Association for Computational Linguistics.

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 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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).

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

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 *Association for Computational Linguistics Workshop*, pages 74–81.

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.

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu. 2021. Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. 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.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

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.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

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.

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.

Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. 2022. Guiding visual question generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1640–1654, Seattle, United States. 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 562–581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Hongwei Zeng, Bifan Wei, Jun Liu, and Weiping Fu. 2023. Synthesize, prompt and transfer: Zero-shot conversational question generation with pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8989–9010, Toronto, Canada. 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.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.

A Annotation Details

A.1 Annotation Instructions

The generated questions are scored on a scale of 1 to 3 on each dimension, detailed instructions for each score are shown in Table 7.

A.2 Annotation Interface

The annotation interface is presented in Figure 3. In the annotation process, annotators should first carefully read the content of the given passage, answer, and question, and then select a score for each dimension.

B Implementation Details of QG models

QG models based on open-source language models are implemented using Hugging Face Transformers, while QG models based on closed-source language models utilize the official open API provided by the respective model. Detailed task instructions we applied for each QG model are presented in Table 8.

Specifically, under the fine-tuning and LoRA settings, we trained QG models separately for each base dataset. In the fine-tuning setting, for the SQuAD dataset, we utilized public fine-tuned models from Huggingface[3], while for the HotpotQA dataset, we conducted our own model fine-tuning as there were few fine-tuned models publicly available. We set the learning rate as 1e-4, warmup steps 500, weight decay 0.01, and the max train epochs as 10 and trained the QG models on a single RTX 3090 GPU. When applying LoRA, we set the learning rate as 1e-4, weight decay as 0.01, and max train epochs as 3 and trained models on an A800 GPU. In few-shot learning, we randomly select 8 examples to provide for the models as recommended in (Min et al., 2022) that model performance does not increase much as the number of examples increases when it reaches 8.

C Data Statistics

We show some statistics of QGEval and comparison with existing benchmarks in Table 9. Compared to existing benchmarks, QGEval covers a

[3]https://huggingface.co/lmqg

Dimensions	Instructions
Fluency	1: The question is incoherent, with imprecise wording or significant grammatical errors, making it difficult to comprehend its meaning. 2: The question is slightly incoherent or contains minor grammatical errors, but it does not hinder the understanding of the question's meaning. 3: The question is fluent and grammatically correct.
Clarity	1: The question is too broad or expressed in a confusing manner, making it difficult to understand or leading to ambiguity. *Particularly, if the generated sentence is not a question but a declarative sentence, it should be considered in this situation.* 2: The question is not expressed very clearly and specifically, but it is possible to infer the question's meaning based on the given passage. 3: The question is clear and specific, without any ambiguity.
Conciseness	1: The question contains too much redundant information, making it difficult to understand its intent. 2: The question includes some redundant information, but it does not impact the understanding of its meaning. 3: The question is concise and does not contain any unnecessary information.
Relevance	1: The question is completely unrelated to the passage. 2: The question is somewhat related to the passage and it asks for non-crucial information related to the passage. 3: The question is relevant to the context, and the information it seeks is crucial to the passage.
Consistency	1: The question contains factual contradictions with the passage or logical errors. 2: The information sought in the question is not fully described in the passage. 3: The information in the question is entirely consistent with the passage.
Answerability	1: The question cannot be answered based on the provided passage. 2: The question can be partially answered based on the provided passage, or the answer to the question can be inferred to some extent. 3: The question can be answered definitively based on the given passage.
Answer Consistency	1: The question cannot be answered by the provided answer. 2: The question can be partially answered using the provided answer. 3: The question can be answered directly using the provided answer.

Table 7: Annotation instructions of evaluation dimensions.

broader range of dimensions, providing a more comprehensive evaluation of generated questions. Additionally, QGEval utilizes a greater variety of models, offering a more robust and thorough assessment and comparison of current QG models.

D Automatic Metrics

Detailed descriptions of the automatic metrics we evaluate are listed as follows:

- **BLEU:** (Papineni et al., 2002), a metric that measures the number of overlapping n-grams between the generated text and a set of gold reference texts.
- **ROUGE:** (Lin, 2004), a recall-oriented metric specifically focuses on the longest common subsequence (LCS) between the generated and reference texts.
- **METEOR:** (Banerjee and Lavie, 2005), a metric computes an alignment between generated texts and reference texts based on the harmonic mean of unigram precision and recall.
- **MoverScore:** (Zhao et al., 2019), a metric measures the Earth Mover's Distance (Kusner et al., 2015) between the distributions of words in the generated text and the reference text.
- **BERTScore:** (Zhang* et al., 2020), a metric computes the semantic similarity of the generated text and reference text by leveraging contextual embeddings from BERT (Devlin et al., 2019).
- **BLEURT:** (Sellam et al., 2020), a learned metric leveraging BERT architecture to evaluate text generation.
- **Q-Metric:** (Nema and Khapra, 2018), a specialized metric designed for the QG task, which considers not only n-gram similarity but also the answerability of questions.
- **QSTS:** (Gollapalli and Ng, 2022), a metric that utilizes the questions' types, entities, and semantic features to evaluate the similarity between questions.

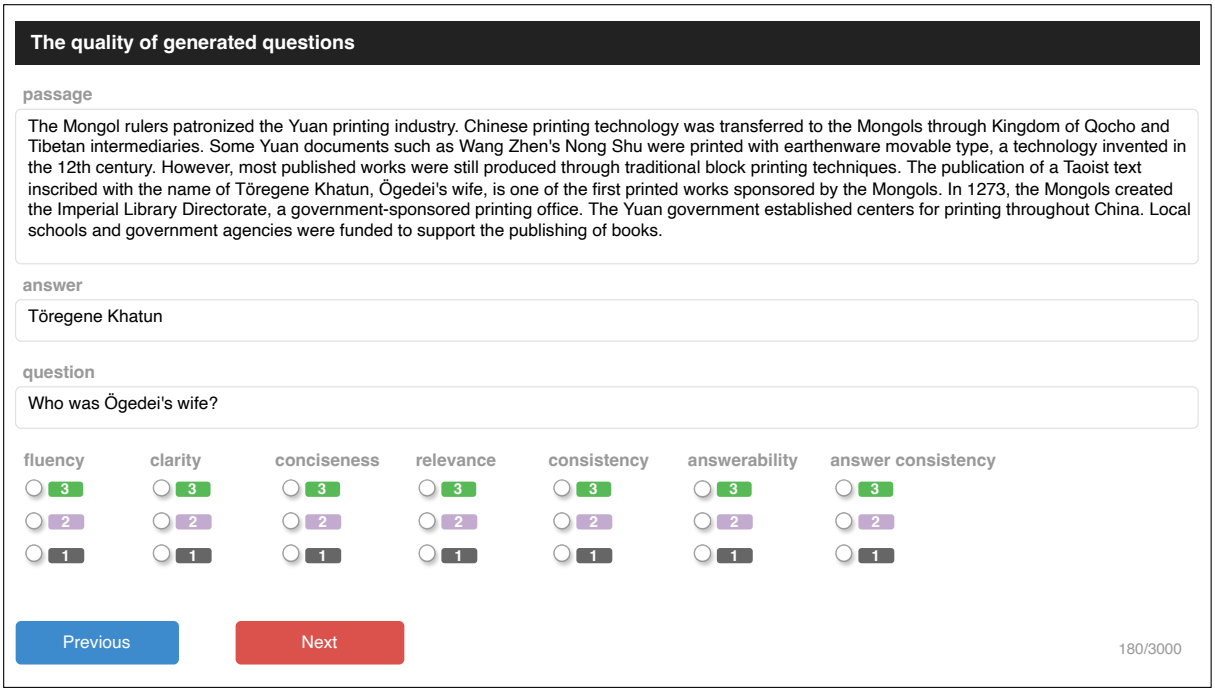


Figure 3: Annotation interface.

- **BARTScore:** (Yuan et al., 2021), a method that formulates evaluating generated text as a text generation task based on the BART model.

- **GPTScore:** (Fu et al., 2023), a framework that leverages the capabilities of generative pre-trained models for evaluation. The intuition of it is similar to BARTScore.

- **UniEval:** (Zhong et al., 2022), a comprehensive framework for evaluating the generated text from multiple explainable dimensions (e.g., fluency) based on T5.

- **QRelScore:** (Wang et al., 2022), a context-aware evaluation method designed for QG, incorporating word-level hierarchical matching based on BERT and sentence-level prompt-based generation techniques based on GPT-2 (Radford et al., 2019).

- **RQUGE:** (Mohammadshahi et al., 2023), a reference-free metric that assesses the answerability of questions based on the QA model's ability to generate an answer to this question within a given context.

We implement the above automatic metrics based on public tools or codes. Specifically, we utilize NLTK[4] to calculate the BLEU and METEOR metrics. As for ROUGE[5], MoverScore[6], and BERTScore[7], we implement them with the corresponding Python packages. For the other metrics, we use their publicly available codes.

E More Experimental Results

E.1 Error Analysis of Generated Questions

We sampled 100 questions generated by QG models and conducted a pilot experiment to analyze the types of errors that occur in these questions. Out of these 100 questions, almost half (42%) contain some degree of error. We find that the generated questions may: 1) be invalid questions, which are declarative sentences or incomplete; 2) be incorrectly phrased; 3) be ambiguously expressed; 4) contain unnecessary copies from the passage that hamper their conciseness; 5) contain inconsistent information with the passage; 6) ask for information not mentioned in the passage, resulting unanswerable based on the passage; 7) do not match with the answers. We present the proportion of each error type among the questions that contain errors in Table 10 and show examples in Table 11.

[4] https://www.nltk.org/
[5] https://pypi.org/project/rouge-score/
[6] https://pypi.org/project/moverscore/
[7] https://pypi.org/project/bert-score/

Models	Instructions
BART-base-finetune BART-large-finetune	{answer} </s> {passage}
T5-base-finetune T5-large-finetune	answer: {answer} context: {passage}
Flan-T5-base-finetune Flan-T5-large-finetune Flan-T5-XL-LoRA Flan-T5-XXL-LoRA	Generate a question based on the given answer and context. Answer: {answer} Context: {passage}
Flan-T5-XL-fewshot Flan-T5-XXL-fewshot	Generate a question based on the given passage and answer. Answer: {example_answer} Context: {example_passage} Question: {example_question} ... Answer: {answer} Context: {passage} Question:
GPT-3.5-turbo-zeroshot GPT-4-zeroshot	Generate a question based on the given answer and context, the generated question must be answered by the given answer. Answer: {answer} Context: {passage} Question:
GPT-3.5-turbo-fewshot GPT-4-fewshot	Generate a question based on the given answer and context, the generated question must be answered by the given answer. Examples: Answer: {example_answer} Context: {example_passage} Question: {example_question} ... Answer: {answer} Context: {passage} Question:

Table 8: Task instructions for different QG models.

Name	#Q	#P	#M	Dimensions	Score Scale	Base dataset
Q-metric (Nema and Khapra, 2018)	3000	—	0	Answerability	1-5	SQuAD, VQA, WikiMovies
SimQG (Gollapalli and Ng, 2022)	500	483	3	Fluency, Relevance, Answerability, Similarity	0-1	SQuAD
Quiz Design Task (Laban et al., 2022)	3164*	7	7	Acceptance	0 or 1	SQuAD
QGEval	3000	200	15	Fluency, Clarity, Conciseness, Relevance, Consistency, Answerability, Answer Consistency	1-3	SQuAD, HotpotQA

Table 9: Detail statistics of QGEval with other benchmarks. #Q: The number of generated questions; #P: The number of passages; #M: The number of QG models. *The number of questions in the Quiz Design Task does not exclude annotations from different annotators.

Error Type	Percentage
Invalid Question	2.38%
Incorrectly Phrased	7.14%
Ambiguous	30.95%
Unnecessary Copy from Passage	16.67%
Inconsistent with Passage	4.76%
Information beyond Passage	19.05%
Mismatch with Answer	47.62%

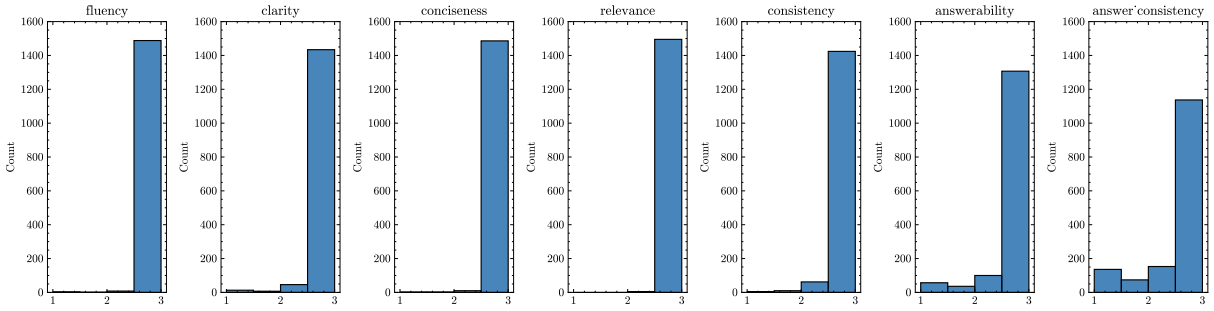
Table 10: Proportion of error types. One question may contain multiple types of errors.

E.2 Annotation Distributions on Different Base Datasets

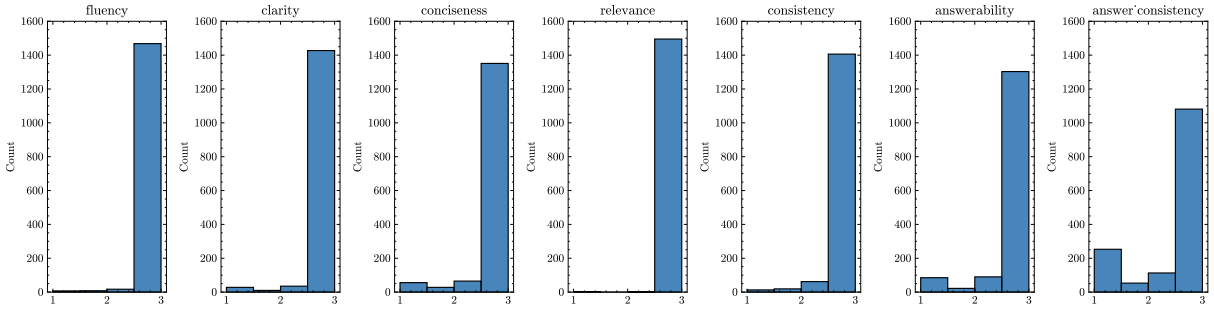
For further insights, we also show the annotation score distribution over each dimension on the SQuAD and HotpotQA datasets in Figure 4a and Figure 4b respectively. We find that compared to questions generated based on SQuAD, questions generated from HotpotQA are more likely to exhibit issues in dimensions such as conciseness, answerability, and answer consistency. This tendency may arise from the fact that reference questions in HotpotQA are predominantly multi-hop questions, resulting in longer question lengths compared to those in SQuAD and posing greater difficulty in terms of answerability.

E.3 Distributions of Automatic Metrics

In Figure 5, we have a look at the distributions of automatic metrics under different human evaluation scores. Taking fluency (linguistic dimension) and answer consistency (task-oriented dimension) as



(a) Annotation score distributions over seven dimensions on SQuAD.



(b) Annotation score distributions over seven dimensions on HotpotQA.

Figure 4: Annotation score distributions over seven dimensions.

examples, we show the distributions of the two automatic metrics that are most and least relevant to the human evaluation results. To better illustrate the distribution results in Figure 5, we round the human scores to the nearest integer, resulting in values of 1, 2, or 3, and then recompute the Pearson Correlations between the two automatic metrics and human scores (i.e., r in the y-axis label).

From the figure, we observe that metrics with low correlations to human scores (e.g., QSTS on fluency and BARTScore-src on answer consistency) cannot accurately score candidates with different human scores. Metrics that achieve higher correlations with human scores (e.g., UniEval on fluency and RQUGE on answer consistency) can correctly assign high scores to high-quality questions (human scores of 3), but they fail to distinguish accurately between questions of lower quality (human scores of 1 or 2).

E.4 Automatic Metrics Used for Ranking QG Models

To further analyze the discriminative ability of automatic metrics across different QG models, we present the average scores of these metrics for questions generated by each model in Table 12. We find that reference-based metrics appear to prefer models based on supervised training since such models excel at generating questions that are similar to the references. This type of metric faces limitations in accurately evaluating questions that are different from references, which makes them struggle to provide precise rankings of the performance of different QG models.

Reference-free metrics address the above limitations of reference-based metrics. The top three models selected based on the scores provided by these metrics partially overlap with those identified through human average scores. However, they also have constraints that result in less precise comparisons of QG models: 1) All of these metrics fail to assign high scores to the reference questions, which is a notable deficiency. 2) Metrics leveraging the generative capabilities of language models appear to exhibit a preference for questions generated by the specific model they utilize. For instance, models with the three highest GPTScore-src scores are the Flan-T5 series (Flan-T5-XL and Flan-T5-XXL), while GPTScore-src also utilizes Flan-T5-XXL as its base model. 3) Metrics designed for specific dimensions are inappropriate for overall performance comparisons across different models. For example, RQUGE is ill-suited for accurately evaluating the overall performance of QG models since it focuses only on the dimensions related to answers.

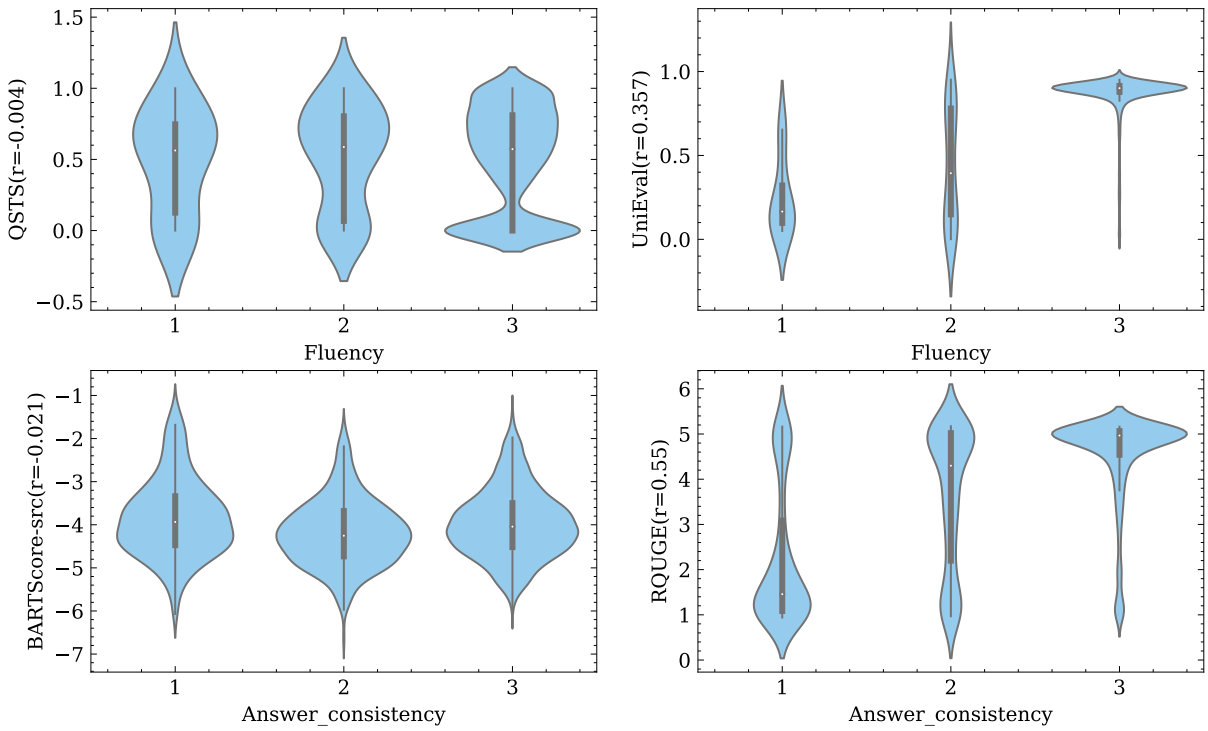


Figure 5: Distributions of automatic metrics under different human scores (1,2,3) on fluency and answer consistency.

Error Type	Example
Invalid Question	**Passage:** *Graptopetalum (leatherpetal) is a plant genus of the family "Crassulaceae". They are perennial succulent plants .....* **Answer:** *yes* **Question:** *Answer: no*
Incorrectly Phrased	**Passage:** *..... The publication of a Taoist text inscribed with the name of Töregene Khatun, Ögedei's wife, is one of the first printed works sponsored by the Mongols.....* **Answer:** *Töregene Khatun* **Question:** *Who was a Taoist text inscribed with the name of gedei's wife?*
Ambiguous	**Passage:** *Although most are non-aligned, some of the best known independent schools also belong to the large, long-established religious foundations, such as the Anglican Church, Uniting Church and Presbyterian Church, ...* **Answer:** *Presbyterian Church* **Question:** *What is another large religious foundation that some of the best known independent schools belong to?*
Unnecessary Copy from Passage	**Passage:** *American burlesque is a genre of variety show. Derived from elements of Victorian burlesque, music hall and minstrel shows, burlesque shows in America became popular in the 1860s and evolved to feature ribald comedy (lewd jokes) and female striptease.* *Brian Newman (born June 10, 1981) is an American jazz musician, singer, and trumpet player. Newman currently holds a residency at the Rose Bar at the Gramercy Park Hotel in New York City and is married to American burlesque performer Angie Pontani.* **Answer:** *American burlesque* **Question:** *Brian Newman (born June 10, 1981) is an American jazz musician, singer, and trumpet player, Newman currently holds a residency at the Rose Bar at the Gramercy Park Hotel in New York City and is married to Angie Pontani, a performer of which genre of variety show, that became popular in the 1860s and evolved to feature ribald comedy (lewd jokes) and female striptease?*
Inconsistent with Passage	**Passage:** *The United States's Sculpin nuclear test series was a group of 7 nuclear tests conducted in 1990-1991. These tests followed the "Operation Aqueduct" series and preceded the "Operation Julin" series.* *The United States's Julin nuclear test series was a group of 7 nuclear tests conducted in 1991–1992. These tests followed the "Operation Sculpin" series, and were the last before negotiations began for the Comprehensive Test Ban Treaty.* **Answer:** *Operation Aqueduct* **Question:** *The United States's Sculpin nuclear test series was a group of 7 nuclear tests conducted in 1990-1991, these tests followed which series, and were the last before negotiations began for the Comprehensive Test Ban Treaty?*
Information beyond Passage	**Passage:** *According to PolitiFact the top 400 richest Americans "have more wealth than half of all Americans combined." .....* **Answer:** *400* **Question:** *According to PolitiFact, who are the richest Americans?*
Mismatch with Answer	**Passage:** *As of August 2013, Skateboardeßr Magazine is primarily a digital skateboarding publication.....its Editor/Photo Editor is Jaime Owens, while the magazine's Publisher is Jamey Stone. On August 19, 2013, the magazine's owner GrindMedia announced that the publication would cease production on October 15, 2013.....* **Answer:** *October 15, 2013* **Question:** *What is the name of the person who is the editor of Skateboarder Magazine?*

Table 11: Examples of errors in generated questions. Errors within questions are highlighted with underlines.

(a) Average scores from reference-based automatic metrics. The three highest scores for each metric are bolded. Abbreviations are as follows. B4:BLEU-4; RL:ROUGE-L; MR:METEOR; BERT:BERTScore; Mover:MoverScore; BRT:BLEURT; $BART_{ref}$:BARTScore-ref; GPT_{ref}:GPTScore-ref; QB4:Q-BLEU4.

Models	B4	RL	MR	BERT	Mover	BRT	$BART_{ref}$	GPT_{ref}	QB4	QSTS
Reference	**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**0.979**	**-2.005**	**-0.385**	**1.000**	**1.000**
BART-base-finetune	0.162	0.444	0.428	0.913	0.638	0.566	-3.502	-1.858	0.374	0.508
BART-large-finetune	0.147	0.427	0.420	0.908	0.630	0.554	-3.640	-1.978	0.361	0.502
T5-base-finetune	0.168	0.467	0.428	0.914	0.642	0.559	-3.480	-1.891	0.374	0.504
T5-large-finetune	**0.177**	**0.491**	**0.446**	**0.918**	**0.652**	**0.583**	**-3.404**	**-1.800**	**0.396**	**0.530**
Flan-T5-base-finetune	**0.171**	0.474	0.438	0.914	0.640	0.566	-3.517	-1.889	0.377	0.513
Flan-T5-large-finetune	0.169	**0.482**	**0.447**	**0.917**	0.639	0.572	-3.432	-1.824	**0.390**	**0.528**
Flan-T5-XL-LoRA	0.160	0.458	0.429	0.911	0.637	0.557	-3.560	-1.907	0.370	0.507
Flan-T5-XXL-LoRA	0.169	0.474	0.427	**0.917**	**0.647**	**0.577**	**-3.409**	**-1.787**	0.384	0.509
Flan-T5-XL-fewshot	0.098	0.380	0.300	0.900	0.609	0.477	-3.731	-2.014	0.280	0.326
Flan-T5-XXL-fewshot	0.110	0.397	0.315	0.906	0.615	0.497	-3.662	-1.943	0.315	0.396
GPT-3.5-turbo-fewshot	0.084	0.330	0.307	0.891	0.592	0.474	-4.033	-2.134	0.247	0.383
GPT-4-fewshot	0.078	0.333	0.340	0.890	0.585	0.491	-4.083	-2.228	0.243	0.457
GPT-3.5-turbo-zeroshot	0.076	0.315	0.295	0.890	0.587	0.471	-4.039	-2.208	0.228	0.353
GPT-4-zeroshot	0.067	0.305	0.323	0.890	0.578	0.490	-4.119	-2.195	0.226	0.430

(b) Average scores from reference-free automatic metrics. The three highest scores for each metric are bolded. Abbreviations are as follows. $BART_{src}$:BARTScore-src; GPT_{src}:GPTScore-src.

Models	$BART_{src}$	GPT_{src}	UniEval	QRelScore	RQUGE
Reference	-4.558	-1.184	0.881	0.045	4.140
BART-base-finetune	-4.011	-0.621	0.901	0.130	4.261
BART-large-finetune	-3.868	-0.632	0.893	**0.155**	**4.363**
T5-base-finetune	-4.041	-0.616	0.892	0.118	4.273
T5-large-finetune	-4.045	-0.601	0.909	0.118	**4.354**
Flan-T5-base-finetune	-4.038	-0.595	0.899	0.120	4.261
Flan-T5-large-finetune	-4.025	-0.603	0.908	0.118	**4.355**
Flan-T5-XL-LoRA	-3.962	**-0.579**	0.900	0.137	4.268
Flan-T5-XXL-LoRA	-4.137	**-0.566**	0.901	0.104	4.289
Flan-T5-XL-fewshot	-4.271	-0.710	0.902	0.071	3.574
Flan-T5-XXL-fewshot	-4.260	**-0.441**	0.904	0.078	3.955
GPT-3.5-turbo-fewshot	**-3.652**	-0.684	0.908	**0.166**	3.555
GPT-4-fewshot	**-3.548**	-0.907	**0.938**	**0.144**	4.240
GPT-3.5-turbo-zeroshot	**-3.652**	-0.869	**0.922**	0.137	3.389
GPT-4-zeroshot	**-3.638**	-1.078	**0.931**	0.121	4.216

Table 12: Average scores of automatic metrics for questions generated by each model.