Evaluation of Question Generation Needs More References

Shinhyeok Oh* **Hyojun Go*** **Yunsung Lee** **Hyeongdon Moon**
Myeongho Jeong **Hyun Seung Lee** **Seungtaek Choi†**
Riiid AI Research
{shinhyeok.oh, hyojun.go, seungtaek.choi}@riiid.co,

Abstract

Question generation (QG) is the task of generating a valid and fluent question based on a given context and the target answer. According to various purposes, even given the same context, instructors can ask questions about different concepts, and even the same concept can be written in different ways. However, the evaluation for QG usually depends on single reference-based similarity metrics, such as n-gram-based metric or learned metric, which is not sufficient to fully evaluate the potential of QG methods. To this end, we propose to paraphrase the reference question for a more robust QG evaluation. Using large language models such as GPT-3, we created semantically and syntactically diverse questions, then adopt the simple aggregation of the popular evaluation metrics as the final scores. Through our experiments, we found that using multiple (pseudo) references is more effective for QG evaluation while showing a higher correlation with human evaluations than evaluation with a single reference.

1 Introduction

Question generation (QG) is the task of generating questions that are relevant to and answerable by given text. Since QG can be applied in not only educational scenarios (Kurdi et al., 2020; Steuer et al., 2021; Moon et al., 2022) but also improving question-answering tasks (Chen et al., 2021; Wang et al., 2018; Yu et al., 2020), designing better QG frameworks and their automatic evaluations have gained more attention (Chakrabarty et al., 2022; Ushio et al., 2022).

However, previous QG works mostly evaluate their methods based on how similar the generated questions are to the gold reference questions (Chan and Fan, 2019; Zhou et al., 2017; Du and Cardie, 2018), using n-gram-based similarity metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Given a single reference, these metrics do not account for the lexical and semantic diversity of questions (Zhang et al., 2020), showing poor correlation with human judgment (Liu et al., 2016; Novikova et al., 2017; Chaganty et al., 2018). Though prior works studied alternative metrics of leveraging language models, such as BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020), such metrics are limited in that the diversity of gold questions is only implicitly represented in the embedding space, rather than data space (or, raw questions).

To explicitly compare with the diverse gold questions in the data space, we propose to augment the single reference question for evaluating QG frameworks, which we call Multi-Reference Evaluation (MRE), by leveraging the few-shot ability of large language models (LLMs) like GPT-3 (Brown et al., 2020) and ChatGPT (OpenAI, 2022). Though there have been efforts to augment references for improving evaluations, they are either limited in other text generation tasks, such as machine translation (Bawden et al., 2020) and question answering (Liu et al., 2021), or the methods are hard to be applied in question generation tasks, as naive LLMs generate some negative (toxic or erroneous) questions (Wang et al., 2022b). Therefore, we utilize LLMs for paraphrasing to augment a reference question, rather than generating new questions from the given context. To the best of our knowledge, we are the first to apply reference augmentation to evaluate the QG frameworks. We briefly summarize our main contributions as follows:

- We propose to augment the single reference for multiple reference evaluation (MRE) that can explicitly consider syntactic and semantic variations of questions. Experimental results on quiz design dataset (Laban et al., 2022) show that the performance of existing metrics can be considerably improved when MRE is

* Equal Contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6358–6367
July 9-14, 2023

applied.

- MRE is metric-agnostic, such that various metrics can be improved with our method. Since each existing metric can discover different insights, such as BLEU for lexical similarity and BERTScore for semantic similarity, MRE can improve these multiple various lenses for investigating QG frameworks.

- We release the augmented reference questions as supplementary materials, which provide an opportunity to reproduce our results for further research. We further validated whether the augmented references are correct or not by human annotators.

2 Methodology

2.1 Single Reference Evaluation (SRE)

Previous works for QG evaluation measure the quality of a generated question q^g in regards to a gold reference question q^r as $M(q^g, q^r)$, where M denotes a similarity metric that is widely used in QG evaluation such as BLEU and ROUGE-L. However, since these metrics suppose only one gold reference, even an appropriate question can be assigned a low score, namely *false positive* problem.

2.2 Multi-Reference Evaluation (MRE)

To deal with this problem, we propose the multi-reference evaluation, where the candidate question q^g is compared with multiple references $Q = \{q_0^r, q_1^r, \ldots, q_N^r\}$:

$$s = \max_i M(q_i^r, q^g) \quad \text{for} \quad i = 0, \ldots, N. \quad (1)$$

By comparing more diverse gold questions with existing metrics, we can measure the more realistic ability of QG frameworks. Note that, as our method could adopt any similarity-based metrics, we can better gain useful insights from various metrics showing different characteristics of generated questions.

However, as it is impractical to collect such multiple references with human annotators, we leverage the recent large language models, specifically GPT-3 and ChatGPT, such that replace Q with $\hat{Q}$. Given a reference question q_0^r, we augment it with N questions:

$$\hat{Q} = \text{LLM}(q_0^r). \quad (2)$$

Note that we give a gold question q_0^r only, rather than the pair of context and question as in (Liu et al., 2021). It is because the zero-shot QG ability of LLMs is reportedly risky for educational purposes (Wang et al., 2022b). We thus use LLMs as a paraphrase generator, which reportedly works well since there is a high correlation between paraphrasing and training paradigms about LLM (Chen et al., 2022).

As GPT-3 is inferior to ChatGPT in the zero-shot settings, here we employ the in-context learning ability of GPT-3, where we give three ChatGPT-paraphrased questions questions as a demonstration for GPT-3 like Appendix A. We will further investigate the correctness of the paraphrased questions in experiments (Section 3.6).

3 Experiments

3.1 Dataset and Evaluation

To verify the effectiveness of MRE, we use quiz design dataset (Laban et al., 2022) for measuring the correlation between automatic question evaluation and human annotation. The quiz design dataset includes 3,164 human-annotated samples, which consist of context, answer, and automatically generated questions. For each sample, the human annotates whether the question is fluent, able to derive the given answer, and fits the context (1) or not (0).

We define the gold human score of a question as the average of the discrete human annotations in $[0, 1]$. Then, we select questions with a human score of 1 as the reference question for the given passage. Finally, for the remaining questions, we measure the Pearson correlation coefficient (Freedman et al., 2007) and Spearman's rank correlation coefficient (Zar, 2005) between the human score and automatic evaluation scores.

3.2 Metrics

Here, as we aim to enhance the existing QG evaluation metrics with multi-reference evaluation, we choose widely used metrics to apply multi-reference evaluation. We apply multi-reference evaluation to BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020). Also, we add RQUGE (Mohammadshahi et al., 2022), which is a reference-free QG evaluation metric, as our baseline. We briefly summarize the metrics used in

	Pearson Correlation					Spearman Correlation				
	SRE	MRE				SRE	MRE			
		HRQ-VAE	GPT-3 (0-shot)	GPT-3 (3-shot)	ChatGPT (0-shot)		HRQ-VAE	GPT-3 (0-shot)	GPT-3 (3-shot)	ChatGPT (0-shot)
BLEU-4	0.2028	0.2443	0.2782	0.3162	**0.3630**	0.2772	0.3224	0.2688	0.3021	**0.3340**
ROUGE-L	0.2908	0.3325	0.3241	0.3447	**0.3799**	0.2787	0.3270	0.3050	0.3330	**0.3637**
RQUGE	0.2932	-	-	-	-	0.2571	-	-	-	-
METEOR	0.3447	0.2968	0.3480	0.3877	**0.4116**	0.3111	0.2822	0.3159	0.3562	**0.3780**
BERTScore	0.3556	0.3634	0.3552	0.3877	**0.4033**	0.3462	0.3568	0.3327	0.3723	**0.3859**
MoverScore	0.4383	0.3835	0.4297	0.4693	**0.4953**	0.3882	0.3643	0.3885	0.4214	**0.4292**
BLEURT	0.4739	0.4287	0.4656	0.4803	**0.5019**	0.4566	0.4193	0.4456	0.4648	**0.4816**

Table 1: Results of the correlation coefficient between measured metrics and human score. The best scores in methodology are in bold, and the best scores in metrics are underlined. These depend on the types of correlation measures. '-' denotes unreported results.

our experiments as follows:

- **BLEU-4** (Papineni et al., 2002) is a metric that utilizes n-gram precision to evaluate the similarity between a generated text and a reference text. The metric counts the number of occurrences of unigrams, bigrams, trigrams, and four-grams that match their corresponding counterparts in the reference text.

- **ROUGE-L** (Lin, 2004) is a metric that utilizes unigram recall to evaluate the similarity between a generated text and a reference text. The metric counts the length of the longest common subsequence as the numerator rather than the exact number of matches.

- **RQUGE** (Mohammadshahi et al., 2022) first predicts answer span with question answering model then computes score with scorer module from given generated question, gold answer, and context. Since RQUGE does not depend on a reference question for evaluation, we only report the correlation of the original RQUGE.

- **METEOR** (Banerjee and Lavie, 2005) measures a score by using a combination of unigram-precision, unigram-recall, and fragmentation measures.

- **BERTScore** (Zhang et al., 2020) utilize contextual embeddings for compute token similarity. We report BERTScore based on `roberta-large`.

- **BLEURT** (Sellam et al., 2020) is a trained metric using a regression model trained on rating data. It combine expressivity and robustness by pre-training a fully learned metric on large amounts of synthetic data, before fine-tuning it on human ratings.

Model	Same answer	Same meaning
GPT-3 (0-shot)	0.77	0.79
GPT-3 (3-shot)	0.83	0.83
ChatGPT (0-shot)	0.92	0.93

Table 2: Human evaluation results of whether paraphrased question by the LLM has the same correct answer and meaning as the reference question.

3.3 Implementation details

We implemented the paraphrasing frameworks by using two LLMs: OpenAI GPT-3 API (Brown et al., 2020) and ChatGPT Webservice (OpenAI, 2022). For GPT-3, we set the model as "text-davinci-003" and the temperature as 0.5. For ChatGPT, we utilized the default setting since we cannot control it. Our prompts are described in Appendix A. We made 20 examples by using LLMs. For additional comparisons with the fine-tuned paraphrasing model, we also implemented HRQ-VAE (Hosking et al., 2022).

3.4 Main Results

As shown in Table 1, we empirically validate the following observations of the advantages of diversified multi-reference evaluation: 1) Our multi-reference evaluation tends to improve the correlation between human score and evaluation metrics. 2) On LLMs, correlation with the human score is high in the order of ChatGPT (0-shot), GPT-3 (3-shot), and GPT-3 (0-shot) paraphrasing framework. Specifically, GPT-3 (3-shot) and ChatGPT paraphrasing framework considerably improve both Pearson correlation and Spearman correlation for all metrics, while paraphrasing with GPT-3 (0-shot)

	$\Delta(MRE - SRE)$					
Human Score	BLEU-4	ROUGE-L	METEOR	BERTScore	MoverScore	BLEURT
1	+ 0.2267	+ 0.1221	+ 0.1034	+ 0.0592	+ 0.0439	+ 0.0400
0	+ 0.0350	+ 0.0846	+ 0.0941	+ 0.0398	+ 0.0190	+ 0.0373

Table 3: Score changes with multiple reference evaluation $\Delta(MRE - SRE)$ through ChatGPT for questions of human score 0 and 1.

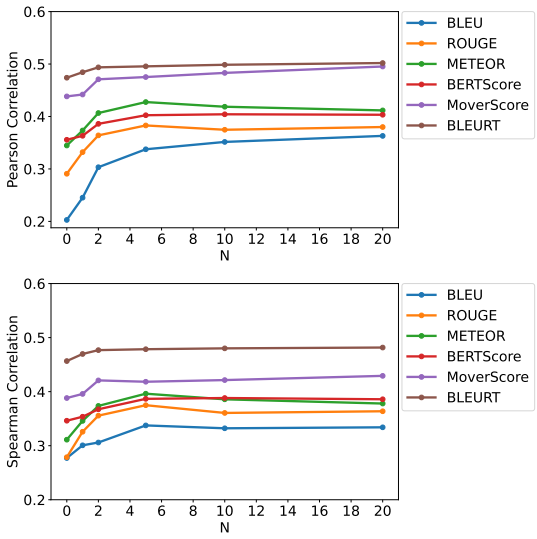

Figure 1: Changes of Pearson and Spearman Correlation coefficients on the number of references generated by ChatGPT (0-shot).

and HRQ-VAE failed at increasing correlations of some metrics.

Also, the increase in correlation through MRE is related to the performance of the paraphrasing framework. As shown in Table 2, the paraphrase of the reference question is better in the order of ChatGPT, GPT-3 (3-shot), and GPT-3 (0-shot). Considering the effect of MRE is also in the same order, we conjecture that the performance of the paraphrasing framework is also important for the effect of MRE. More details in Table 2 are described in Section 3.6.

3.5 Analysis for MRE

The effect of N We analyze the effect of the number of reference questions N by changing N to 1, 2, 5, 10, and 20. Figure 1 shows the change of the correlation coefficient according to the change of N. The results show that even if only one augmented reference question is used, the correlation is higher than that of the single reference evaluation. Also, if more augmented reference questions are used, the correlation with the human score increases and becomes saturated when N exceeds a certain level ($N \approx 5$).

Score change with multi-reference evaluation We further explore how MRE changes original metrics. Specifically, we report average score differences between the original metric and the multi-reference version of it with ChatGPT for accepted and unaccepted candidate questions. Questions with the human score of 1 and 0 are considered accepted questions and unaccepted questions, respectively.

As shown in Table 3, multi-reference evaluation increases the score of accepted questions relatively more than that of an unaccepted question. For example, BLEU-4 score increases by 0.2267 for accepted questions, compared to 0.0350 for unaccepted questions. These results mean that multi-reference evaluation makes original metrics more correlated with the human score by enlarging the score of acceptable questions than unacceptable questions.

3.6 Human Evaluation of Question Paraphrase

The assumption of multi-reference evaluation is that most paraphrased questions with LLMs can serve the meaning like the gold questions. We conduct a human study to validate this assumption. For each of GPT-3 (0-shot), GPT-3 (3-shot), and ChatGPT, we sample 50 pairs of reference questions and paraphrased questions and annotate each pair whether the paraphrased questions have the same meaning and have the same answer compared to reference questions. Specifically, we ask two annotators to evaluate with a binary rating (1 for "same" and 0 for "not same"). As shown in Table 2, 92% and 93% of the questions paraphrased by ChatGPT are evaluated as having the same answer and meaning, respectively. In addition, even when paraphrasing with GPT3 3-shot, it has the same meaning and the same answer at a high rate. We refer to Appendix B for more details about human annotation.

	Generated Question	Approach	Reference Question	B-4	R-L	BS	BR	Human
E1	What is the definition of sustainable energy?	SRE	What does it mean if energy is sustainable?	0.00	0.27	0.68	0.75	1.00
		MRE-B-4	What is the definition of sustainable energy?	1.00	-			
		MRE-R-L	What is the definition of sustainable energy?	-	1.00			
		MRE-BS	What is the definition of sustainable energy?	-	-	1.00	-	
		MRE-BR	What is the definition of sustainable energy?	-	-	-	0.97	
E2	What are some examples of renewable energy sources?	SRE	What are some renewable energy sources?	0.00	0.86	0.87	0.83	1.00
		MRE-B-4	What are some examples of renewable energy?	0.53	-			
		MRE-R-L	What are some examples of alternative energy sources?	-	0.87			
		MRE-BS	What are some examples of renewable energy?	-	-	0.95	-	
		MRE-BR	What are some examples of renewable energy?	-	-	-	0.85	
E3	How is energy sustainable?	SRE	What does it mean if energy is sustainable?	0.00	0.33	0.74	0.77	0.00
		MRE-B-4	What does sustainable energy mean?	0.00	-			
		MRE-R-L	What does it mean if energy is sustainable?	-	0.33			
		MRE-BS	What does sustainable energy mean?	-	-	0.76	-	
		MRE-BR	What does it mean if energy is sustainable?	-	-	-	0.77	

Table 4: Examples of SRE and MRE results. MRE-B-4, MRE-R-L, MRE-BS, and MRE-BR denotes to use BLEU-4, ROUGE-L, BERTScore, and BLEURT as M, respectively. Reference Question for SRE represents the given reference question q_0^r, and the Reference Question for MRE-B-4, MRE-R-L, MRE-BS, and MRE-BR represent one of $\hat{Q}$ that obtained the max score for each measure.

3.7 Case Study

For example in E1 in Table 4, one of the texts in paraphrased references matches the generated question. MRE achieves gains over SRE by 1.00 ($0.00 \rightarrow 1.00$) on BLEU-4, and we found a positive effect on all other metrics. In E2, the text that received the highest score among paraphrased references differs from each metric. We can observe that MRE works well by showing that you can choose one of the paraphrased references that are measured to be similar for each metric. Moreover, score increases suggest that MRE leads to positive shifts in the metric scores when the human score is 1 (E1, E2). However, the score to utilize MRE cannot be lower than SRE in any example because MRE takes the maximum score for the true reference and paraphrased references. Thus, if the human score is low, it is important to have a small negative effect. One may ask about the risk of MRE giving a higher score than SRE for wrong questions as in E3. However, we argue that it doesn't weaken the strength of MRE as the gaps between SRE and MRE for wrong questions are relatively smaller than that for correct questions, which we compared in Table 3.

4 Conclusion & Future Work

In this paper, we studied the problem of evaluating the question generation frameworks, and observed that automatically augmenting the reference question with large language models is surprisingly effective, showing higher correlations with human-annotated scores. Though we evaluated the effectiveness of multiple reference evaluations for test-time evaluations, where the gold human score is given, we hope future research to explore other scenarios, such as measuring validation performance (asking how much the test performance can be actually improved) and multi-reference training as in (Jeong et al., 2021). Exploring other tasks (machine translation and document summarization) or generation methods (giving context and the reference question together to LLMs) would be interesting for future research.

5 Limitations

Inapplicability to reference-free evaluation: Since our MRE supposes that there is an available reference question to be augmented (paraphrased), it is not applicable to reference-free question evaluations such as QRelScore (Wang et al., 2022a) and RQUGE (Mohammadshahi et al., 2022).

Inapplicability for answer-unconditional QG frameworks: MRE can't be applied to answer-unconditional QG frameworks because it only augments the reference question by paraphrasing without considering other possible questions of supposing other answers.

Large computations: To generate multi-reference questions, our method requires inference of large language models, which results in huge computational costs. Therefore, this can become burdensome as the test dataset grows.

6 Ethical Considerations

We honor and support the ACL code of Ethics. In order to conduct our human annotation for paraphrased sentences, two humans are recruited. We make sure that humans would be paid a wage of 15 dollars per hour.

References

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

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Tuhin Chakrabarty, Justin Lewis, and Smaranda Muresan. 2022. Consistent: Open-ended question generation from news articles.

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

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.

Zheng Chen, Hu Yuan, and Jiankun Ren. 2022. Zero-shot domain paraphrase with unaligned pre-trained language models. *Complex & Intelligent Systems*, pages 1–14.

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.

David Freedman, Robert Pisani, Roger Purves, and Ani Adhikari. 2007. Statistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2489–2501, Dublin, Ireland. Association for Computational Linguistics.

Myeongho Jeong, Seungtaek Choi, Jinyoung Yeo, and Seung-won Hwang. 2021. Label and context augmentation for response selection at dstc8. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2541–2550.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

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 North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ruibo Liu, Jason Wei, and Soroush Vosoughi. 2021. Language model augmented relevance score. 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 6677–6690, Online. Association for Computational Linguistics.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.

Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. Evaluating the knowledge dependency of questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10512–10526, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

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.

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.

Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. 2021. I do not understand what i cannot define: Automatic question generation with pedagogically-driven content selection. *arXiv preprint arXiv:2110.04123*.

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

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022a. Qrelscore: Better evaluating generated questions with deeper understanding of context-aware relevance. *arXiv preprint arXiv:2204.13921*.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2193–2203.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022b. Towards human-like educational question generation with large language models. In *Artificial Intelligence in Education*, pages 153–166, Cham. Springer International Publishing.

Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si. 2020. Review-based question generation with adaptive instance transfer and augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 280–290.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 662–671. Springer.

A LLM prompts

In this section, we describe the details of how to paraphrase the questions with LLMs. To avoid confusion, we report it with an example question.

A.1 ChatGPT (0-shot) & GPT-3 (0-shot)

We use 0-shot input for ChatGPT and GPT-3 as:

```
Please paraphrase the following sentence 20 times:
Where is the Statue of Liberty located?
```

A.2 GPT-3 (3-shot)

For the 3-shot prompt, we use the examples paraphrased by ChatGPT as the 3-shot. Our 3-shot prompt is described as:

```
Please paraphrase the following sentence 20
↪  times:

Sentence: What is this software found useful for?
1. For what purpose is this software deemed
↪  useful?
2. What are the uses of this software?
3. Can you tell me what this software is useful
↪  for?
4. What are the benefits of this software?
5. In what ways is this software found to be
↪  useful?
6. What are the applications of this software?
7. Can you explain the usefulness of this
↪  software?
8. What does this software excel at?
9. Can you tell me what tasks this software is
↪  useful for?
10. In what scenario is this software useful?
11. Can you describe the utility of this
↪  software?
12. What is the purpose of this software?
13. For what is this software commonly used?
14. What does this software aid in?
15. Can you tell me the function of this
↪  software?
16. What are the advantages of this software?
17. What is this software good for?
18. In what ways does this software provide
↪  value?
19. What is the benefit of using this software?
20. Can you tell me what this software is
↪  commonly used for?

Sentence: Who is the girl?
1. Can you tell me the name of the young lady?
2. Who is the female in question?
3. Could you identify the girl for me?
4. To whom are you referring as the girl?
5. The girl, who is she?
6. Could you tell me who the girl is?
7. Who is the young woman being spoken of?
8. Can you name the girl in question?
9. Who is the lady in question?
10. Could you give me the name of the girl?
11. Who is being referred to as the girl?
12. Can you tell me who the female is?
13. Who is the girl being discussed?
14. Can you identify the young lady?
15. The girl, can you tell me her name?
16. Who is the subject of the girl?
17. Can you name the female in question?
18. Who is the girl you are asking about?
19. Can you provide the name of the girl?
20. Who is the young woman being referred to?

Sentence: Sentence: Where is the Eiffel Tower?
1. Can you tell me the location of the Eiffel
↪  Tower?
2. Could you inform me where the Eiffel Tower is
↪  situated?
3. I'm wondering where the Eiffel Tower is
↪  located?
4. The Eiffel Tower, where can I find it?
5. Could you give me the whereabouts of the
↪  Eiffel Tower?
6. The Eiffel Tower, where is it located?
7. Can you indicate the location of the Eiffel
↪  Tower?
8. Can you provide me with the location of the
↪  Eiffel Tower?
9. Where can I find the Eiffel Tower?
10. The Eiffel Tower, where is it situated?
11. Can you tell me where the Eiffel Tower is
↪  located?
12. Could you give me the location of the Eiffel
↪  Tower?
13. Where is the Eiffel Tower situated?
14. The Eiffel Tower, where is it found?
15. Could you inform me where the Eiffel Tower
↪  can be found?
16. Can you give me the whereabouts of the Eiffel
↪  Tower?
17. Where is the Eiffel Tower located?
18. The Eiffel Tower, where is it positioned?
19. Can you indicate the whereabouts of the
↪  Eiffel Tower?
20. Can you provide me with the whereabouts of
↪  the Eiffel Tower?

Sentence:
```

B Human Annotation

Two annotators participate in our study. All the pairs from paraphrasing LLMs are randomly shuffled and anonymized, and each pair is evaluated by the following two dimensions:

Same Answer Human annotators check whether the paraphrased question has the same answer as the reference question. Annotation is performed by binary rate, 1 for "having the same answer" and 0 for "having the different answer".

Same meaning It checks whether the paraphrased question has the same meaning as the reference question. Humans annotate the question as 1 for "having the same meaning" and 0 for "having a different meaning". The inter-annotator agreement is 0.24 for the same meaning, and 0.21 for the same answer. Although the agreement was low due to the difference in their standards, the model preference was clearly preserved for both annotators.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☑ A2. Did you discuss any potential risks of your work?
5

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly, correct grammar for all sections

B ☑ Did you use or create scientific artifacts?

3.1

☑ B1. Did you cite the creators of artifacts you used?
3.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ 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.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
just using api service for augmentation

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.3

☑ 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?
3

☑ 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.)?
3.2, 3.3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3.6

☐ 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.?
Not applicable. Left blank.

☑ 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)?
6

☒ 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?
just evaluation for automatic generated data

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
just evaluation for automatic generated data