Generative Interpretation: Toward Human-Like Evaluation for Educational Question-Answer Pair Generation

Hyeonseok Moon[1], **Jaewook Lee**[1], **Sugyeong Eo**[1]
Chanjun Park[2], **Jaehyung Seo**[1] **and Heuiseok Lim**[1†]
[1]Department of Computer Science and Engineering, Korea University
[2]Upstage
[1]{glee889,jaewook133,djtnrud,seojae777,limhseok}@korea.ac.kr
[2]chanjun.park@upstage.ai

Abstract

Educational question-answer generation has been extensively researched owing to its practical applicability. However, we have identified a persistent challenge concerning the evaluation of such systems. Existing evaluation methods often fail to produce objective results and instead exhibit a bias towards favoring high similarity to the ground-truth question-answer pairs. In this study, we demonstrate that these evaluation methods yield low human alignment and propose an alternative approach called Generative Interpretation (GI) to achieve more objective evaluations. Through experimental analysis, we reveal that **GI** outperforms existing evaluation methods in terms of human alignment, and even shows comparable performance with GPT3.5, only with BART-large.

1 Introduction

Asking questions about the passage enhances children's literacy development (Blewitt et al., 2009; Sim and Berthelsen, 2014). In the context of children's learning, educational question-answer generation (QAG) has gained considerable attention due to its practical utility (Xu et al., 2022; Dugan et al., 2022; Yao et al., 2022). QAG frameworks aim to generate relevant question-answer (QA) pairs based on a given story passage. With the significant research focus on QAG, numerous frameworks have been proposed to generate diverse and accurate QA pairs (Lee et al., 2020; Johnson et al., 2022; Eo et al., 2023)

While the generation capability of QAG has witnessed significant advancements, precise automatic evaluation remains a challenge. Current automatic evaluation metrics for QAG primarily rely on assessing textual similarity, such as ROUGE(Lin, 2004), and BERTscore(Zhang et al.), with respect to the ground-truth(GT) QA pairs (Dugan et al., 2022; Yao et al., 2022). However, we have observed that GT similarity seldom poses high score for the high relevancy to the given passage, but only prefer GT similar QA pair, which follows misalignment with human assessment (Graham, 2015).

We consider ***evaluation*** to be a crucial factor in education of QAG, as inaccurate assessments can result in improper guidance (Shanmugavelu et al., 2020). Considering the role of QAG in the educational field, automatic evaluation methods serve as substitutes for human judgment that discriminate the most appropriate QA pair for the given passage. In such setting, an improper evaluation approach may restrict creative responses (Bullough Jr, 1992) and skew the purpose of the education towards mimicking answers from the GT QA dataset.

In an effort to mitigate such limitations, we propose a more objective and precise evaluation method, Generative Interpretation (**GI**). **GI** employs a generative QAG model trained with GT QA pairs and selectively measures teacher-forced logits that are highly relevant in evaluating QA pairs. By evaluating each QA pair in a reference-free manner, **GI** enables even objective assessment that cannot be figured out via comparison between GT QA pairs. We figure out that **GI** can yield even higher human correlation, compared with the existing evaluation method. In particular, we demonstrate that only with utilizing the BART-large model structure (Lewis et al., 2020), **GI** can offer comparable performance to the ChatGPT (GPT3.5) evaluation (OpenAI-Blog, 2022).

2 Related Works

QAG frameworks aim to generate numerous QA pairs by given a passage (Xu et al., 2022; Liu et al., 2020; Jerome et al., 2021). In considering diversity in QA even enhances children's intellectual and literacy development (Dillon, 2006; Shanmugavelu et al., 2020), current QAG studies mainly focus on enhancing diversity of the generating QA pairs (Yao et al., 2022; Zhao et al., 2022; Eo et al., 2023),

Findings of the Association for Computational Linguistics: EACL 2024, pages 2185–2196
March 17-22, 2024 ©2024 Association for Computational Linguistics

without harming relevancy to the given passage (Dugan et al., 2022; Lee et al., 2020). However, such methods only adopt the GT-similarity based evaluation method, which can yield biased results toward GT similar QA pairs (Graham, 2015).

3 Preliminary

The evaluation on the QAG framework is performed by measuring the quality of the candidate QA set $C = \{(q^c_j, a^c_j)\}^{N_c}_{j=1}$, generated by the QAG framework given a passage P. Existing methods measure the textual similarity between the C and the GT QA set $R = \{(q^r_i, a^r_i)\}^{N_r}_{i=1}$. We denote the textual similarity metric as **Metric**, where existing studies primarily adopt two measures, ROUGE and BERTscore. Considering multi-reference and multi-candidate setting, we can find two strategies in evaluating C.

Concat-Metric For the comprehensive evaluation, Zhao et al. (2022) concatenates all the QA pairs in a single sequence for each QA pair set, and estimates **Metric** between them. In this case, estimated quality of C, denoted as $s_{\texttt{Concat}}$, can be computed as equation (1). We denote $[\cdots]$ as a sequentialized concatenation of all elements.

$$\begin{gathered} r_i = [\ q^r_i\ \ a^r_i\ ],\ r = [\ r_1, \cdots, r_{n_r}\ ] \\ c_j = [\ q^c_j\ \ a^c_j\ ],\ c = [\ c_1, \cdots, c_{n_c}\ ] \\ s_{\texttt{Concat}} = \textbf{Metric}\,(r, c) \end{gathered} \tag{1}$$

MAP@N-Metric Yao et al. (2022); Eo et al. (2023); Xu et al. (2022) find the most similar QA pair in C, for each QA pair in R[1]. In other words, we calculate the highest **Metric** for each QA pair in R, that can be derived by comparison with any QA pair in C. We can compute the estimated quality of C, denoted as $s_{\texttt{MAP}}$, as shown in equation (2).

$$\begin{gathered} \text{metric}_{i,j} = \textbf{Metric}([q^r_i\ \ a^r_i], [q^c_j\ \ a^c_j]) \\ s_{\texttt{MAP}} = \frac{1}{N_r} \sum_{i=1}^{N_r} \max_j \{\text{metric}_{i,j}\}^{N_c}_{j=1} \end{gathered} \tag{2}$$

Challenges in Evaluation In applying human evaluation, QAG systems are generally estimated by the following aspects (Dugan et al., 2022; Eo et al., 2023; Zhao et al., 2022): (i) **Relevancy to the passage** that determines whether the QA pair is relevant to the passage, (ii) **Answerability of the answer** that shows whether the answer can be regarded as an appropriate response to the question, and (iii) **Grammatical plausibility** of the generated QA pair. However, we argue that the existing automatic evaluation method of measuring similarity to GT has limitations in satisfying the above requirements and only evaluates whether the QA is similar to GT without evaluating the objective quality of the QA.

4 Generative Interpretation (GI)

GI estimates the adequacy of the generated QA pair, which encompasses relevancy to the passage and the connectivity between the QA. Similar with BARTscore (Yuan et al., 2021), we adopt QA generation model and take teacher-forced logits of the QA generation. In particular, we train the QA generation model θ to return concatenated sequence of the QA pair, by feeding passage and the question start tokens, with $\mathcal{L}_{CE}$ shown in equation (3).

$$\mathcal{L}_{CE} = -\frac{1}{N_r} \sum_{i=1}^{N_r} \prod_{l=1}^{N_{r_i}} \mathbf{P}_\theta(r_{i,l} | r_{i,<l}, q^r_{i,<\mathbf{n_s}}, P) \tag{3}$$

The number of question start tokens are priorly set by a hyper-parameter $\mathbf{n_s}$. We feed start tokens of each question as a part of input sequence, to alleviate the question type bias[2]. In utilizing θ, we can estimate **GI** as follows:

4.1 Teacher-Forced Inference

GI is estimated by the teacher-forced logits of the candidate QA pair, calculated by θ. Precisely, we denote the probability of l^{th} token in $c_j = [c_{j,1}, \cdots, c_{j,n_c}]$, to be generated by θ as $prob^c_{j,l}$. Then we calculate the score for C, $\boldsymbol{s}_{\mathbf{GI}}$, as follows:

$$prob^c_{j,l} = \mathbf{P}_\theta(c_{j,l} | c_{j,<l}, q^c_{j,<\mathbf{n_s}}, P) \tag{4}$$

$$\boldsymbol{s}_{\mathbf{GI}} = \frac{1}{N_c} \sum_{j=1}^{N_c} \left[\frac{1}{N_{c_j}} \sum_{l=2}^{N_{c_j}-1} prob^c_{j,l} \right] \tag{5}$$

[1] Xu et al. (2022) inversely matched the most appropriate reference for each candidate. However, as noted in Yao et al. (2022) and Eo et al. (2023), we find that such setting may bear unfair results, and set the baseline as in Yao et al. (2022).

[2] We argue that, in estimating the relevancy of the question to the given passage, question types that generally determined by the preceding tokens of the question should not be considered. For instance, "why" question can be generated for any passage. In this regard, we hypothesize that relevancy of the question to the passage is only determined by the proceeding sequences

CLASS	Evaluation Method	Rel P-QA	Rel Q-A	Rel Avg	Usb	Rdb	Overall Avg
(a) GT Similarity	MAP@N-**ROUGE**	0.25708	0.28495	0.27101	0.44938	0.27702	0.31711
	Concat-**ROUGE**	0.23987	0.29014	0.26501	0.44401	0.27455	0.31214
	MAP@N-**BS**	0.31421	0.34464	0.32943	0.45315	0.33133	0.36083
	Concat-**BS**	0.30326	0.31240	0.30783	0.44174	0.33771	0.34878
(b) ChatGPT	GPT3.5 (P)	0.71699	0.36462	0.54080	0.13104	**0.43409**	0.41168
	GPT3.5 (QA)	0.73321	0.44916	0.59118	0.36482	0.35204	0.47481
	GPT4 (P)	0.70115	0.50391	0.60253	0.41532	0.14611	0.44162
	GPT4 (QA)	**0.78532**	**0.64633**	**0.71583**	**0.47354**	0.39561	**0.57520**
(c) **GI**	**GI** - T5	0.63169	0.41141	0.52155	0.32029	0.42928	0.44817
	GI - BART	**0.64525**	**0.46689**	**0.55607**	**0.40833**	**0.44438**	**0.49121**
	GI $_{SS}$ - T5	0.28245	0.23667	0.25956	0.25465	0.29370	0.26687
	GI $_{SS}$ - BART	0.17870	0.19743	0.18806	0.15051	0.33684	0.21587

Table 1: Experimental results in the respect of the human correlation (pearson-r). We denote **BS** as BERTscore, (P) as content-wise evaluation, (QA) as QA-wise evaluation, and **Rel Avg** as the average of the **Rel P-QA** and **Rel Q-A**. In estimating **GI** , we set $\mathbf{n}_s$ as 4.

GI works as a reference-free evaluation method, that can evaluate any QA pairs GT-independently. In particular, logits of question position in c_j determines the relevancy of the QA pair to the given passage, and answer position in c_j reflects the answerability of answer in c_j. Additionally, as logit reflects generation possibility, we can also judge the readability of QA pair via **GI** .

Unlike in training phase, **GI** is calculated as the mean of probabilities to prevent probability deterioration led by a single outlier. Also, the probability at the [BOS] and [EOS] position are excluded from the calculation for mitigating unintended bias.

4.2 Syntactic Similarity with Inference Output

The high performance of **GI** may solely contributed to the vast linguistic capability of θ. For clarifying the validity of **GI** , we establish another baseline evaluation method, **GI** $_{SS}$, that estimates the textual similarity between the generation output of θ with C. By comparing **GI** with **GI** $_{SS}$, we verify the effectiveness of **GI** in evaluating QAG, with relieved dependency on QAG model capacity. More details are described in Appendix F

5 Experiments

5.1 Experimental Settings

We adopt Fairytale QA dataset (Xu et al., 2022) in our experiments, as we find it as the most appropriate dataset fitted in educational purpose and is constructed by the human experts. We proceed human evaluation on QA pairs for 20 passages, generated by four QAG systems, including gold QA pair. As in Eo et al. (2023), we evaluate four aspects: **Relevancy P-QA (Rel P-QA)** that estimates the relevance between QA pairs and a passage, **Relevancy Q-A (Rel Q-A)** that evaluates whether a question and its corresponding answer are correctly matched, **Usability (Usb)** estimating practical usability of the QA pair in educational field, **Readability (Rdb)** that indicates grammatical correctness. More precise details are dealt with Appendix A.

Adequacy for each metric is estimated by the pearson-r and kendall-tau correlation with human evaluation score (Koo and Li, 2016). Main results report pearson-r results (Freitag et al., 2021), and kendall-tau is dealt in the Appendix D. We adopt two pretrained language models in establishing **GI** : T5(Raffel et al., 2020) and BART(Lewis et al., 2020), and measure ROUGE-L F1-score in estimating ROUGE, and F1-score for BERTscore. More extensive details about training and experimental settings are described in Appendix B.

5.2 ChatGPT as an Evalutor

One may wonder that all the evaluation process can be charged to ChatGPT owing to its extraordinarily high performance(Peng et al., 2023; Ouyang et al., 2022). In particular, several other tasks such as essay assessment (Chiang and Lee, 2023; Liu et al., 2023) adopted ChatGPT (OpenAI-Blog, 2022) in evaluation and show high human alignment. Con-

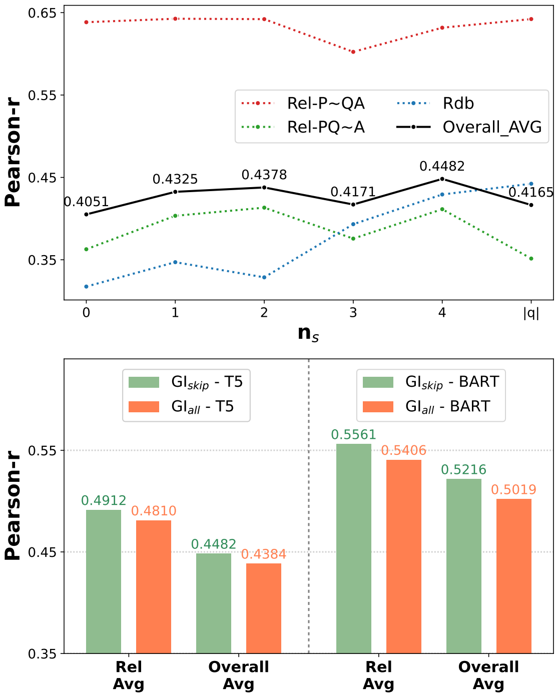


Figure 1: Case studies considering **GI** . Upper figure shows the human alignment variants depending on ($\mathbf{n}_s$), where $|q|$ denotes the full length of question. Below figure demonstrates the effect of the logits in BOS and EOS position, where "skip" denotes the intended **GI** , and "all" considers all the logits, including BOS and EOS position.

sidering these, we check the performance of ChatGPT in evaluation of QAG, and verify the difficulty of evaluation in QAG and the effectiveness of **GI** .

In utilizing ChatGPT, we adopt the prompts and human instructions adopted to the prior studies (Yuan et al., 2022; Eo et al., 2023). As current evaluation protocol encompasses passage-wise evaluation (Eo et al., 2023) and QA-wise evaluation (Dugan et al., 2022; Xu et al., 2022), we experiment both of settings with specialized prompts. More details about the prompt engineering is described in Appendix C.

5.3 Main Results

Estimating similarity between GT may bear suspicious results As shown in class (a) of Table 1, we find that GT similarity based metrics shows even low human alignment especially for the relevancy aspects. This implies that similarity between GT suffers severe challenge in determining whether the QA pair is relevant to the given passage. Rather, it shows unexpectedly high correlation with usability aspect. These results indicates that existing evaluation methods can be regarded as suspicious evaluators.

ChatGPT is a decent evaluator Results in class (b) of Table 1 shows that ChatGPT is a decent evaluator for the QAG. We find that QA-wise evaluation (*i.e.* GPT#(QA)) highly promote the evaluation performance of ChatGPT. Specifically, GPT4 shows the prominent performance, while GPT3.5 demonstrates relatively moderate performance, which implies the difficulty of evaluation for QAG.

GI is a trust-worthy evaluator Considering all the results in Table 1, specifically in the respect of class (C), we find that **GI** shows great human alignment (More details are in Appendix E). **GI** - BART outperforms existing evaluation methods, and even surpasses the performance of GPT3.5.

In particular, while **GI** shows comparable performance with GPT3.5, **GI** $_{SS}$ does not even outperform ROUGE. This result indicate the methodologies applied in **GI** enables more objective and human-like evaluation of each QA pair.

5.4 Case Study

In estimating **GI**, we exclude "question start tokens" by feed it as a input, and dismiss logits in [BOS] and [EOS] positions, considering them as spurious factor that may lead to unintended bias. Figure 1 demonstrates case studies regarding them. We find that adjusting the number of question start tokens ($\mathbf{n}_s$) lead to even higher performance, by dismissing irrelevant logits in evaluating QA pairs. Similarly, we find that logits in [BOS] and [EOS] position also lead to unintended bias and decreases human alignment. More detailed results are described in Appendix D.4.

6 Conclusion

In this study, we focus on challenges in existing evaluation methods of educational QAG that measuring quality based on the similarity with GT QA pairs. We find out that existing automatic evaluation methods show inferior human alignment especially in measuring relevancy to the passage and question-answer pair. As alternatives, we propose more objective evaluation methodology, **GI**, that can relieve several challenges in existing metrics. We shows that **GI** demonstrates even higher human alignment than GPT3.5, only with BART-large. We plan to extend **GI** to more general metric that can cover more generalized question generation tasks.

7 Limitations

We find the effective of **GI** is only verified by the two model structures. We argue that **GI** can be applied to any large language models and more objective evaluation can be exploited by adopting more powerful language models. While this works only deals with BART-large and T5-base models due to the resource limitation, we plan to extend our experiments and urge future studies regarding model extension.

Additionally, we hope to clarify that our human evaluation was conducted with 240 QA pairs. Though it may seem small, we consider it to be a sufficient number for drawing general conclusions as compared to other studies that conducted human evaluations on approximately 100 QA pairs (Dugan et al., 2022). Notably, even on relatively ambiguous evaluation criteria, the achieved Krippendorff's alpha score of 0.59 indicates our results are sufficiently reproducible and reliable.

8 Ethics Statement

We recruited participants by posting an announcement on a university community site that can be viewed by all members of the university; the individuals who participated in the experiment have no relationship with the authors outside of the present study. All the participants were provided with full disclosure about the purpose and process of the experiment before proceeding. We required from them official English proficiency scores (TOEIC, TOEFL), and only invited as evaluators those who had scores equivalent to or higher than 90 out of 100. All of the participants were asked for a B.A degree certificate in Education, ensuring that the evaluators had comparable levels of understanding in English and educational theory. In this process, all personally identifiable information from the human evaluators was immediately discarded after verification.

We paid the evaluators $0.34 per evaluated QA, and we asked each evaluator to conduct a total evaluation on 240 QA pairs. We awarded a week for the evaluation period, and granted them autonomy in setting their own start and end times of evaluation. All evaluations were conducted on identical UI sites and everyone evaluated the same passage and same QA. We clearly state that there were absolutely no ethical issues that could be raised related to the human evaluation.

Acknowledgements

This work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-2020-0-01819). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

References

Pamela Blewitt, Keiran M Rump, Stephanie E Shealy, and Samantha A Cook. 2009. Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology*, 101(2):294–304.

Robert V Bullough Jr. 1992. Beginning teacher curriculum decision making, personal teaching metaphors, and teacher education. *Teaching and Teacher Education*, 8(3):239–252.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

James T Dillon. 2006. Effect of questions in education and other enterprises. In *Rethinking schooling*, pages 145–174. Routledge.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-unaware question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926.

Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee, Changwoo Chun, Sungsoo Park, and Heuiseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Yvette Graham. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.

Bill Jerome, Rachel Van Campenhout, and Benny G Johnson. 2021. Automatic question generation and the smartstart application. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pages 365–366.

Benny G Johnson, Jeffrey S Dittel, Rachel Van Campenhout, and Bill Jerome. 2022. Discrimination of automatically generated questions used as formative practice. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 325–329.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Raymond G Miltenberger. 1990. Assessment of treatment acceptability: A review of the literature. *Topics in Early Childhood Special Education*, 10(3):24–38.

OpenAI. 2023. Gpt-4 technical report.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ganesan Shanmugavelu, Khairi Ariffin, Manimaran Vadivelu, Zulkufli Mahayudin, and Malar Arasi RK Sundaram. 2020. Questioning techniques and teachers' role in the classroom. *Shanlax International Journal of Education*, 8(4):45–49.

Susan Sim and Donna Berthelsen. 2014. Shared book reading by parents with young children: Evidence-based practice. *Australasian Journal of Early Childhood*, 39(1):50–55.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is ai's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744.

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

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2022. Selecting better samples from pre-trained llms: A case study on question generation. *arXiv preprint arXiv:2209.11000*.

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

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5073–5085.

A Dataset Details

For constructing our test dataset for the verification, we randomly extract 20 passages from Fairytale QA test dataset. Then we adopt three QAG systems: Yao et al. (2022)(FQAG), (Dugan et al., 2022)(SQG) and Eo et al. (2023)(DQAG). Then we generate three QA pairs for each passage with each QAG system. Additionally, we select three QA pairs from QA pair set that linked to the corresponding passage.

Subsequently, we proceeded human evaluation for each systems, including GT QA pairs. Human evaluation processes are the same as Eo et al. (2023). All human evaluators hold a bachelor's degree in education. We assess the following four aspects estimating the quality of the QA pairs.

- **Relevancy P-QA**: This evaluates the relevance between a passage and a QA pair. If either question or answer is not relevant, it is irrelevant.
- **Relevancy Q-A**: This evaluates whether a question and its corresponding answer are correctly generated. If either of them is awkward, it is considered.
- **Usability**: This evaluates whether the generated QA pairs can be used for education purposes.
- **Readability**: This evaluates whether the generated QA pairs are grammatically right.

In this study, we revised notation utilized in Eo et al. (2023), for allievating confusion with the educational domains (Miltenberger, 1990). We amend the term "Acceptability" to "Relevancy Q-A", and subsequently replace the term "Relevancy" to "Relevancy P-QA". We got 0.5900 krippendorff's alpha score over all the human evaluation results, and obtained the maximum score for the **Relevancy Q-A** (0.6355) (Krippendorff, 2011).

B Training Details

For implementing **GI** , we adopt BART-large and T5-base model structure provided by the Huggingface(Wolf et al., 2020) framework (under Apache License 2.0). In training models for **GI** , we utilize a single RTX A6000 GPU. Each training is proceeded with AdamW optimizer(Loshchilov and Hutter, 2017) with learning rate $1e-04$ and batch size 32. We select the best performing model among different learning rate settings: $\{2e-04, 1e-04, 3e-05, 1e-05\}$.

C ChatGPT Details

For the implementation of ChatGPT in our experiments, we utilize GPT-3.5(gpt-3.5-turbo-0301) (Ouyang et al., 2022) and GPT-4(gpt-4-0314) (OpenAI, 2023) and applied 0.7 temperature. We establish our prompts inspired by the previous works (Yuan et al., 2022), which aims at question generation utilizing LLM. Following Liu et al. (2023) and Mehri and Eskenazi (2020), we compose our prompt to include the human instruction for each aspect. In particular, we construct two types of the prompts that (1) evaluating each QA pair (**QA-wise**), and (2) evaluating QA pairs that correspond to the same passage (**content-wise**).

For **QA-wise**, each evaluation factor was scored on a 5-point Likert scale ranging from 1 to 5, which was then averaged across passages to obtain a final score, and **content-wise** was scored on a scale ranging from 0-3, with 1 point for each of the 3 QA pairs generated from a passage that met the criteria. All scores were re-scaled to values between 0 and 1. Utilized prompts are shown in figure 2 and figure 3.

D Detailed Experimental Results

D.1 Case Studies for ROUGE, BERTscore

Several existing QAG studies report ROUGE-L or BERTscore measured with precision or recall (Yao et al., 2022; Dugan et al., 2022). In this study, we point out that there is no clear standard in selecting one among precision, recall, or F1, and clarify human alignments of these methods in QAG. Experi-

Pearson-r		Rel	Rel	Rel	Usb	Rdb	Overall
		P-QA	Q-A	Avg			Avg
MAP@N **ROUGE**	**P**	0.26977	0.27628	0.27302	0.43007	0.25500	0.30778
	R	0.25773	0.29819	0.27796	0.45150	0.29500	0.32561
	F	0.25708	0.28495	0.27102	0.44938	0.27702	0.31711
MAP@N **BS**	**P**	0.33944	0.32486	0.33215	0.44164	0.32231	0.35706
	R	0.26110	0.34458	0.30284	0.43809	0.32605	0.34245
	F	0.31422	0.34465	0.32943	0.45315	0.33133	0.36084
Concat **ROUGE**	**P**	0.23128	0.28824	0.25976	0.45023	0.27928	0.31226
	R	0.24550	0.28453	0.26502	0.42944	0.25752	0.30425
	F	0.23988	0.29014	0.26501	0.44402	0.27455	0.31215
Concat **BS**	**P**	0.23540	0.27444	0.25492	0.40738	0.33385	0.31277
	R	0.34941	0.32761	0.33851	0.44394	0.31634	0.35933
	F	0.30327	0.31240	0.30783	0.44175	0.33771	0.34878

Table 2: Experimental results on the variants of the existing methods, in the respect of the human correlation (pearson-r). We denote **BS** as BERTscore, **P** as a precision, **R** as a recall, and **F** as a F1-measure.

mental results are described in Table 2 and Table 3.

Experimental results shows that precision can be a more human-like measure compared with F1 measures, depending on its evaluating method. However, we argue that these results still cannot give considerable human alignments compared with **GI**.

D.2 Experimental Results on the Kendall-tau Correlation

In out main results, we only report pearson-r correlation which indicates high correlation. For more objective verification, we additionally implement kendall-tau (Koo and Li, 2016) verification. Experimental results are shown in Table 4. We view the kendall-tau result as an auxiliary indicator as in Freitag et al. (2021).

D.3 Results of GI variants

We report detailed experimental results regarding the Section 5.4. In this section, we demonstrate that discriminating "necessary part" in generating and accumulating logits of θ is the essential part in estimating **GI**.

Table 5 describes the whole results of the experiments on the variant of $\mathbf{n}_\mathrm{s}$. $\mathbf{n}_\mathrm{s}$ determines the extent of the information fed to the generative model θ. Note that θ is supervised to return the generation probability of QA pair. We hypothesize that question start tokens (which can include interrogative) determines the category of the corresponding question (Eo et al., 2023), and hardly related to the relevancy between passage. In this regard, we find that feeding question start tokens to the θ can yield more objective generation probability in judging "whether the question is relevant to the given passage". If $\mathbf{n}_\mathrm{s}$ is zero, generated probability can be influenced by the interrogative distribution of the training data, which may lead to unintended bias in estimating relevancy of the QA pair to the given passage. On the contrary, if $\mathbf{n}_\mathrm{s}$ is equal to the length of question, we find that θ cannot properly identify the relationship between questions and answers, as the whole sequence of question is granted as input. Experimental results on Table 5 support our claims, which demonstrates the best performance when $\mathbf{n}_\mathrm{s}$ is set to 4.

Table 6 implies the reason we established the calculation process of **GI** as in Equation (5). In accumulating teacher-forced logits for calculating **GI**, we exclude probability yielded by decoding `[BOS]` and `[EOS]` positions. Note that the motivation of **GI** is estimating the plausibility of each QA pair, given a corresponding passage. In considering this, we find that probability obtained from `[BOS]` and `[EOS]` positions does not give meaningful information in estimating relevancy. As described in Table 6, we find that by following our intuition, we can enhance human alignment of **GI** (**GI** with **Skip**).

D.4 System-level Evaluation

For better elaborate the practical utility of **GI** , we implement system-level evaluation, and the followings Table 7 reveal our results. For the human

Kendall-τ		**Rel P-QA**	**Rel Q-A**	**Rel Avg**	**Usb**	**Rdb**	**Overall Avg**
MAP@N **ROUGE**	**P**	0.13474	0.18781	0.16128	0.29157	0.17221	0.19658
	R	0.18849	0.23668	0.21259	0.31734	0.22419	0.24167
	F	0.16760	0.20674	0.18717	0.31490	0.20622	0.22386
MAP@N **BS**	**P**	0.16092	0.23053	0.19572	0.29679	0.20360	0.22296
	R	0.23641	0.25479	0.24560	0.27719	0.24994	0.25458
	F	0.20489	0.25895	0.23192	0.29679	0.25601	0.25416
Concat **ROUGE**	**P**	0.18198	0.19887	0.19043	0.30290	0.21283	0.22414
	R	0.11798	0.17614	0.14706	0.26548	0.15151	0.17778
	F	0.15232	0.20648	0.17940	0.29489	0.20203	0.21393
Concat **BS**	**P**	0.19643	0.20733	0.20188	0.27983	0.26216	0.23644
	R	0.23756	0.22908	0.23332	0.29214	0.17221	0.23275
	F	0.20063	0.23891	0.21977	0.30309	0.25832	0.25024

Table 3: Experimental results on the variants of the existing methods, in the respect of the human correlation (kendall-tau). We denote **BS** as BERTscore, **P** as a precision, **R** as a recall, and **F** as a F1-measure.

CLASS	**Evaluation Method**	**Rel P-QA**	**Rel Q-A**	**Rel Avg**	**Usb**	**Rdb**	**Overall Avg**
(b) ChatGPT	GPT3.5 (P)	0.15595	0.21445	0.18520	0.01381	0.30393	0.17204
	GPT3.5 (QA)	0.27470	0.37293	0.32382	0.22972	0.16018	0.25938
	GPT4 (P)	0.13703	0.39058	0.26381	0.36151	0.13887	0.25700
	GPT4 (QA)	0.36983	0.55765	0.46374	0.35242	0.28195	0.39047
(c) **GI**	**GI** - T5	0.18354	0.24220	0.21287	0.15498	0.14992	0.18266
	GI - BART	0.11988	0.31544	0.21766	0.23449	0.16657	0.20910
	GI $_{SS}$ - T5	0.11326	0.15859	0.13592	0.16306	0.19535	0.15756
	GI $_{SS}$ - BART	0.05952	0.11022	0.08487	0.21764	0.20368	0.14776

Table 4: Experimental results in the respect of the human correlation estimated by the kendall-tau coefficient.

Evaluation Method	$\mathbf{n}_s$	**Rel P-QA**	**Rel Q-A**	**Rel Avg**	**Usb**	**Rdb**	**Overall Avg**
GI - BART	**0**	0.57054	**0.50290**	0.53672	**0.47142**	0.33041	0.46882
	1	0.61186	0.42744	0.51965	0.42291	0.31076	0.44324
	2	0.53692	0.45698	0.49695	0.43070	0.38009	0.45117
	3	0.65388	0.44694	0.55041	0.37583	0.40666	0.47083
	4	0.64526	0.46689	**0.55608**	0.40834	**0.44439**	**0.49122**
	|q|	**0.66631**	0.42024	0.54328	0.31224	0.42419	0.45575
GI - T5	**0**	0.63840	0.36281	0.50061	0.30156	0.31755	0.40508
	1	**0.64264**	0.40344	0.52304	0.33688	0.34705	0.43250
	2	0.64213	0.41332	0.52772	**0.36695**	0.32873	0.43778
	3	0.60241	0.37569	0.48905	0.29713	0.39317	0.41710
	4	0.63170	**0.41142**	**0.52156**	0.32030	0.42928	**0.44817**
	|q|	0.64218	0.35147	0.49683	0.23016	**0.44226**	0.41652

Table 5: Experimental results of **GI** on the variants of $\mathbf{n}_s$. We report pearson-r correlation with human evaluation results.

Evaluation Method		Rel P-QA	Rel Q-A	Rel Avg	Usb	Rdb	Overall Avg
GI - BART	**Skip**	0.64526	0.46689	**0.55608**	0.40834	0.44439	**0.49122**
	All	0.61167	0.46945	0.54056	0.45129	0.39171	0.48103
GI - T5	**Skip**	0.63170	0.41142	**0.52156**	0.32030	0.42928	**0.44817**
	All	0.58763	0.41620	0.50192	0.36617	0.38357	0.43839

Table 6: Experimental results of **GI** on the variants of $\mathbf{n}_S$. We report pearson-r correlation with human evaluation results.

	Eo et al. (2023)	Yao et al. (2022)	Dugan et al. (2022)	**Ground-Truth**
Human Evaluation	0.7775	0.7475	0.7483	0.8658
GPT3.5	0.9666	0.8999	0.9124	0.9916
GPT4	0.9874	0.8916	0.9249	1.0000
ROUGE-L	0.3643	0.3567	0.3545	1.0000
BERTscore	0.9790	0.9799	0.9788	1.0000
GI (ours)	0.8903	0.8483	0.7799	0.9163

Table 7: System level evaluation results

evaluation results, we measured average score for the four aspect we guaged (i.e. Rel P-QA, Rel Q-A, Usb, Rdb)

Our experiments reveal that **GI** attains high alignment with human evaluation, and gives informative results. While ROUGE and BERTscore yield mere difference across different systems, **GI** shows distinctive measure. More detailed evaluation results for each datapoint (i.e. evaluation for each QA set) are described in Figure 4.

E Qualitative Analysis

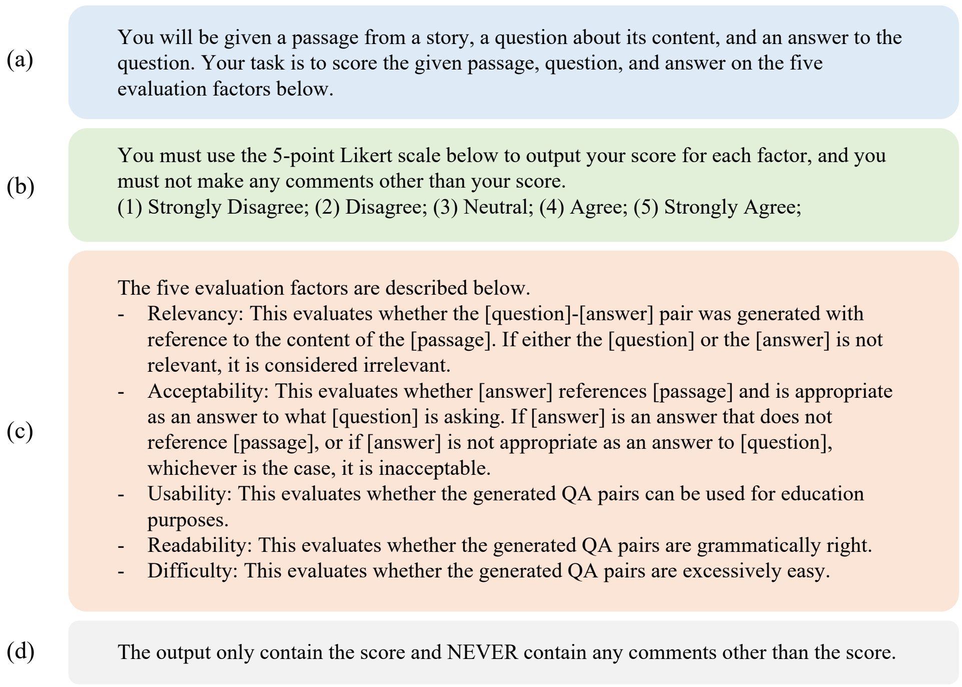
To verify the practical usability of **GI**, we qualitatively analyze evaluation results proceeded in this study. We randomly extract three representative samples from our test dataset. As shown in Figure 4, **GI** shows high human alignment and similar tendency with GPT3.5 and GPT4. However, ROUGE shows even contrary results with human evaluation results, as it only reflects similarity with GT QA pairs. BERTscore provided high score with greater than 0.95 for all the QA pairs, that we can hardly determine which QA pair is decent or not. Our qualitative analysis support our main results, and further implies practical utility of **GI**.

F Detailed Description of GI $_{SS}$

The method GI_{ss}, being experimented for illustration in our proposal, signifies that the effectiveness of GI is not merely reliant on the language understanding capability of the trained model itself. Essentially, the evaluation model θ is trained to generate QAs by taking inputs from the passage and question start tokens.

In utilizing θ, GI is estimated by utilizing logit values. On the other hand, GI_{ss} evaluates the appropriateness of the generated output by comparing it with the candidate QAs.

Consider an evaluation candidate QA set $[(q_1, a_1), \ldots, (q_n, a_n)]$ for passage P. By utilizing the trained model θ, we induce generation of answer $a_i^{'}$ by taking P and each q_i as inputs. Afterwards, we compare the textual similarity (ROUGE-L) between each a_i and $a_i^{'}$. Using this generation-based evaluation method GI_{ss}, we observed significantly lower performance than GI. This experiment can essentially be regarded as a demonstration that, even when using the same evaluation model, the logit-based evaluation method we proposed is even more effective.



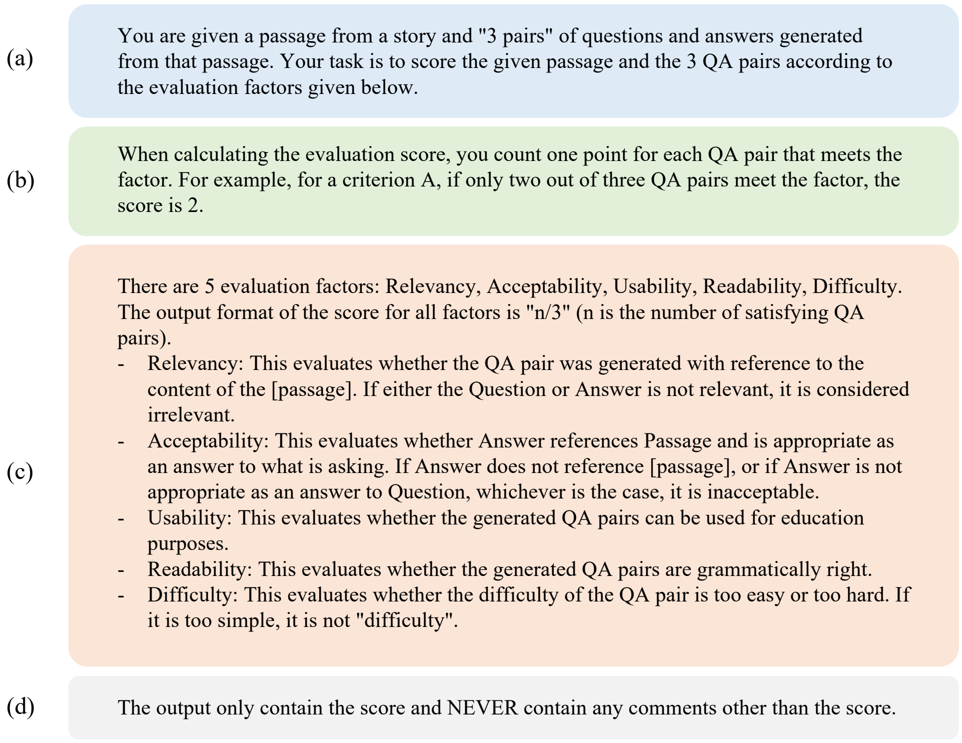
(a) You will be given a passage from a story, a question about its content, and an answer to the question. Your task is to score the given passage, question, and answer on the five evaluation factors below.

(b) You must use the 5-point Likert scale below to output your score for each factor, and you must not make any comments other than your score.
(1) Strongly Disagree; (2) Disagree; (3) Neutral; (4) Agree; (5) Strongly Agree;

(c) The five evaluation factors are described below.
- Relevancy: This evaluates whether the [question]-[answer] pair was generated with reference to the content of the [passage]. If either the [question] or the [answer] is not relevant, it is considered irrelevant.
- Acceptability: This evaluates whether [answer] references [passage] and is appropriate as an answer to what [question] is asking. If [answer] is an answer that does not reference [passage], or if [answer] is not appropriate as an answer to [question], whichever is the case, it is inacceptable.
- Usability: This evaluates whether the generated QA pairs can be used for education purposes.
- Readability: This evaluates whether the generated QA pairs are grammatically right.
- Difficulty: This evaluates whether the generated QA pairs are excessively easy.

(d) The output only contain the score and NEVER contain any comments other than the score.

Figure 2: Prompts utilized in QA-wise evaluation of ChatGPT



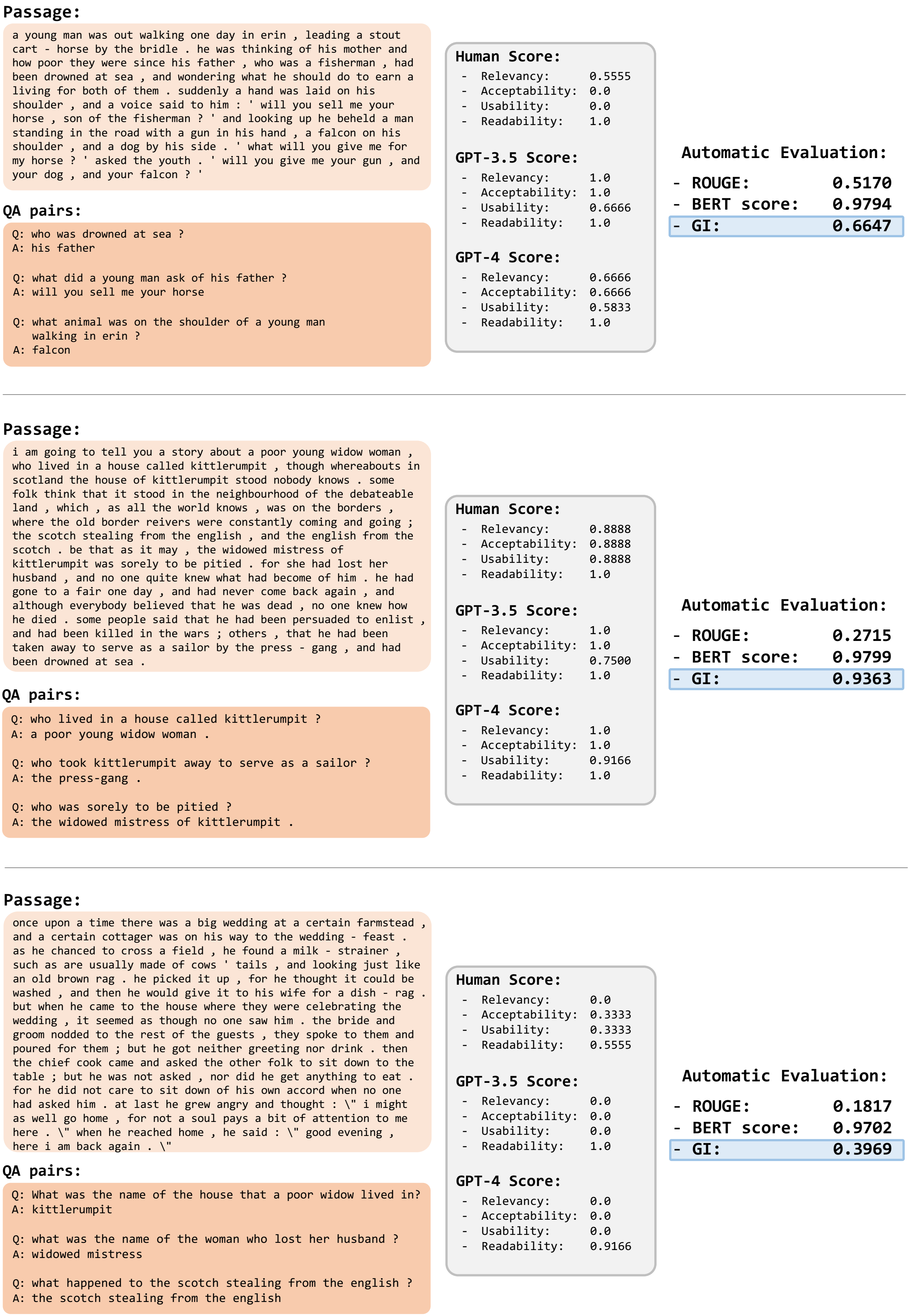
(a) You are given a passage from a story and "3 pairs" of questions and answers generated from that passage. Your task is to score the given passage and the 3 QA pairs according to the evaluation factors given below.

(b) When calculating the evaluation score, you count one point for each QA pair that meets the factor. For example, for a criterion A, if only two out of three QA pairs meet the factor, the score is 2.

(c) There are 5 evaluation factors: Relevancy, Acceptability, Usability, Readability, Difficulty. The output format of the score for all factors is "n/3" (n is the number of satisfying QA pairs).
- Relevancy: This evaluates whether the QA pair was generated with reference to the content of the [passage]. If either the Question or Answer is not relevant, it is considered irrelevant.
- Acceptability: This evaluates whether Answer references Passage and is appropriate as an answer to what is asking. If Answer does not reference [passage], or if Answer is not appropriate as an answer to Question, whichever is the case, it is inacceptable.
- Usability: This evaluates whether the generated QA pairs can be used for education purposes.
- Readability: This evaluates whether the generated QA pairs are grammatically right.
- Difficulty: This evaluates whether the difficulty of the QA pair is too easy or too hard. If it is too simple, it is not "difficulty".

(d) The output only contain the score and NEVER contain any comments other than the score.

Figure 3: Prompts utilized in content-wise evaluation of ChatGPT



Passage:

a young man was out walking one day in erin , leading a stout cart - horse by the bridle . he was thinking of his mother and how poor they were since his father , who was a fisherman , had been drowned at sea , and wondering what he should do to earn a living for both of them . suddenly a hand was laid on his shoulder , and a voice said to him : ' will you sell me your horse , son of the fisherman ? ' and looking up he beheld a man standing in the road with a gun in his hand , a falcon on his shoulder , and a dog by his side . ' what will you give me for my horse ? ' asked the youth . ' will you give me your gun , and your dog , and your falcon ? '

QA pairs:

Q: who was drowned at sea ?
A: his father

Q: what did a young man ask of his father ?
A: will you sell me your horse

Q: what animal was on the shoulder of a young man walking in erin ?
A: falcon

Human Score:

- Relevancy: 0.5555
- Acceptability: 0.0
- Usability: 0.0
- Readability: 1.0

GPT-3.5 Score:

- Relevancy: 1.0
- Acceptability: 1.0
- Usability: 0.6666
- Readability: 1.0

GPT-4 Score:

- Relevancy: 0.6666
- Acceptability: 0.6666
- Usability: 0.5833
- Readability: 1.0

Automatic Evaluation:

- ROUGE: 0.5170
- BERT score: 0.9794
- GI: 0.6647

Passage:

i am going to tell you a story about a poor young widow woman , who lived in a house called kittlerumpit , though whereabouts in scotland the house of kittlerumpit stood nobody knows . some folk think that it stood in the neighbourhood of the debateable land , which , as all the world knows , was on the borders , where the old border reivers were constantly coming and going ; the scotch stealing from the english , and the english from the scotch . be that as it may , the widowed mistress of kittlerumpit was sorely to be pitied . for she had lost her husband , and no one quite knew what had become of him . he had gone to a fair one day , and had never come back again , and although everybody believed that he was dead , no one knew how he died . some people said that he had been persuaded to enlist , and had been killed in the wars ; others , that he had been taken away to serve as a sailor by the press - gang , and had been drowned at sea .

QA pairs:

Q: who lived in a house called kittlerumpit ?
A: a poor young widow woman .

Q: who took kittlerumpit away to serve as a sailor ?
A: the press-gang .

Q: who was sorely to be pitied ?
A: the widowed mistress of kittlerumpit .

Human Score:

- Relevancy: 0.8888
- Acceptability: 0.8888
- Usability: 0.8888
- Readability: 1.0

GPT-3.5 Score:

- Relevancy: 1.0
- Acceptability: 1.0
- Usability: 0.7500
- Readability: 1.0

GPT-4 Score:

- Relevancy: 1.0
- Acceptability: 1.0
- Usability: 0.9166
- Readability: 1.0

Automatic Evaluation:

- ROUGE: 0.2715
- BERT score: 0.9799
- GI: 0.9363

Passage:

once upon a time there was a big wedding at a certain farmstead , and a certain cottager was on his way to the wedding - feast . as he chanced to cross a field , he found a milk - strainer , such as are usually made of cows ' tails , and looking just like an old brown rag . he picked it up , for he thought it could be washed , and then he would give it to his wife for a dish - rag . but when he came to the house where they were celebrating the wedding , it seemed as though no one saw him . the bride and groom nodded to the rest of the guests , they spoke to them and poured for them ; but he got neither greeting nor drink . then the chief cook came and asked the other folk to sit down to the table ; but he was not asked , nor did he get anything to eat . for he did not care to sit down of his own accord when no one had asked him . at last he grew angry and thought : \" i might as well go home , for not a soul pays a bit of attention to me here . \" when he reached home , he said : \" good evening , here i am back again . \"

QA pairs:

Q: What was the name of the house that a poor widow lived in?
A: kittlerumpit

Q: what was the name of the woman who lost her husband ?
A: widowed mistress

Q: what happened to the scotch stealing from the english ?
A: the scotch stealing from the english

Human Score:

- Relevancy: 0.0
- Acceptability: 0.3333
- Usability: 0.3333
- Readability: 0.5555

GPT-3.5 Score:

- Relevancy: 0.0
- Acceptability: 0.0
- Usability: 0.0
- Readability: 1.0

GPT-4 Score:

- Relevancy: 0.0
- Acceptability: 0.0
- Usability: 0.0
- Readability: 0.9166

Automatic Evaluation:

- ROUGE: 0.1817
- BERT score: 0.9702
- GI: 0.3969

Figure 4: Qualitative analysis. ROUGE generally give high score to the GT-similar QA pairs and thereby shows low human alignment. BERTscore typically imposed high score that we can hardly figure out indicator in determining superior QA pair. **GI** shows high human alignment and similar tendency with GPT3.5 and GPT4.