# **Automatic Answerability Evaluation for Question Generation**

# Zifan Wang, Kotaro Funakoshi, Manabu Okumura

Tokyo Institute of Technology {wangzf,funakoshi,oku}@lr.pi.titech.ac.jp

#### **Abstract**

Conventional automatic evaluation metrics, such as BLEU and ROUGE, developed for natural language generation (NLG) tasks, are based on measuring the n-gram overlap between the generated and reference text. These simple metrics may be insufficient for more complex tasks, such as question generation (QG), which requires generating questions that are answerable by the reference answers. Developing a more sophisticated automatic evaluation metric, thus, remains an urgent problem in QG research. This work proposes **PMAN** (**P**rompting-based **M**etric on **AN**swerability), a novel automatic evaluation metric to assess whether the generated questions are answerable by the reference answers for the QG tasks. Extensive experiments demonstrate that its evaluation results are reliable and align with human evaluations. We further apply our metric to evaluate the performance of QG models, which shows that our metric complements conventional metrics. Our implementation of a GPT-based QG model achieves state-of-the-art performance in generating answerable questions.

Keywords: Evaluation Metrics, Chain-of-Thought (CoT) Prompting, ChatGPT

#### 1. Introduction

Question Generation (QG) (Du et al., 2017; Yuan et al., 2017; Zhou et al., 2018) and Multi-hop Question Generation (MQG) (Pan et al., 2020; Su et al., 2020; Fei et al., 2022) are tasks of generating questions that are answerable by the specified reference answers from given passages. Commonly used automatic evaluation metrics for QG tasks are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics measure the n-gram overlap between the generated and reference text. However, it cannot assess whether the generated questions are answerable by the reference answers, though it is an essential requirement in the QG tasks. Consequently, previous work often supplements these metrics with human evaluations to assess the answerability. However, human evaluations have apparent disadvantages: (1) They are time-consuming and expensive, and (2) Different work has conducted human evaluations under varying conditions, making it difficult to compare the performance of different QG models. Therefore, developing an automatic evaluation metric to assess answerability has become a crucial problem.

To address this problem, we propose PMAN, Prompting-based Metric on ANswerability, a novel automatic evaluation metric for the QG tasks, which assesses whether the generated questions are answerable with the supposed answers. PMAN leverages the high instruction-following capabilities of recent large language models (LLMs) such as ChatGPT (OpenAI, 2022) and responds to input with "YES" or "NO" as its answerability.

We conducted extensive experiments to test the

reliability of our proposed metric, using both manually created and model-generated samples. Our experiments demonstrate that the metric assessments are reliable and align with human evaluations. Furthermore, we applied our metric to evaluate the performance of QG models, showing that our metric complements conventional metrics. Additionally, we implemented a GPT-based QG model, which achieves state-of-the-art (SOTA) performance in generating answerable questions. The main contributions of this work are two-fold: (1) Proposing a novel automatic evaluation metric that measures answerability and conducting extensive experiments to validate its reliability; and (2) Demonstrating that our metric complements conventional metrics and a GPT-based MQG model achieves SOTA performance in generating answerable multi-hop questions.

### 2. Related Work

While automatic metrics measuring n-gram overlap are still commonly used for evaluating the performance of NLG models, they have been widely criticized for their low correlation with human judgments. Furthermore, these metrics cannot measure more nuanced properties, such as whether a generated question is answerable by a reference answer. To address these limitations, most work complements the automatic metrics with human evaluations (Xie et al., 2020; Ji et al., 2021), which are time-consuming and expensive.

The only significant work we could find that proposed a metric for measuring answerability is Nema and Khapra (2018). They observed that specific components of a question are more relevant to its answerability, and the overlap of words in these components with words in the gold question

| Α | Your task is to determine if the reference answer delimited by triple dashes is the answer to the question   |
|---|--|
|   | delimited by angle brackets, according to the passage delimited by triple backticks.                         |
| С | To solve the problem do the following:   |
|   | - First, give your own answer to the question.   |
|   | - Then compare your answer to the reference answer delimited by triple dashes and evaluate if the answer     |
|   | delimited by triple dashes is correct or not. Don't decide if the reference answer is correct until you have |
|   | answered the question yourself.  |
| Α | - If the reference answer is correct, respond "YES", otherwise respond "NO".                                 |
|   |  |
| В | Passage: " 'passage context " '  |
|   | Question: <question context=""></question>   |
|   | Reference answer:answer context  |

Figure 1: CoT prompting for PMAN. Section A is the task description. B is the data. C is the CoT part.

should be assigned a higher score than the overlap of random words. For example, if the gold question is "Who was the director of Titanic?", question A, "director of Titanic?" would be assigned a much lower score than question B, "Who was the director of?" by the existing overlap-based metrics. However, most humans would consider sentence A more answerable than B, as it contains a key component, a named entity "Titanic." Nema and Khapra (2018) identified four components (question types, content words, function words, and named entities) and proposed assigning a higher score for their overlap. Nevertheless, this approach is still an overlap-based metric, ignoring the one-to-many nature of the QG tasks. As pointed out by Yuan et al. (2017) and Ji et al. (2022), we can ask diverse questions that are answerable by the same answer. For example, "What person directed Titanic?" would still have the same answer as the gold guestion "Who was the director of Titanic?" They are almost equally answerable. However, no overlap-based metric would assign a reasonable score to such questions.

Recently, new reference-free frameworks have been proposed to automatically evaluate NLG models. (Kocmi and Federmann, 2023; Wang et al., 2023; Liu et al., 2023; Mohammadshahi et al., 2023). They employed large language models to assign scores for specific properties of generated texts, such as coherence and relevance. This framework is reference-free such allows the generation of diverse questions, and can measure even more nuanced properties such as answerability, which were almost impossible to measure using the previous automatic metrics. Especially, Mohammadshahi et al. (2023) also proposed a metric for evaluating the quality of automatically generated questions, and explored the idea of leveraging the LLMs to answer the question. However, our work differs from theirs in the following aspects: (1) Their QA model extracts a span in the input passage as the answer and thus cannot answer questions for which the answer does not appear in the passage (e.g., a question whose an-

swer is either "yes" or "no"). Our model instead generates the answer to avoid this limitation; (2) Their metric requires an appropriate supervised training dataset in the target domain and language for the Span Scorer. Our approach does not; (3) Their metric is not strictly intended to evaluate answerability. They describe their metric as "to compute the acceptability of the candidate", which assesses the overall quality of generated questions, as a replacement for previous metrics. It mixes up multiple aspects including fluency and relevance. Therefore, when the assessed score is not high, it is not clear to the metric user whether it is due to fluency or answerability. Our work instead aims to assess whether the questions can be answered by the reference answer, as a complementary of previous metrics. To our knowledge, it is still the first work to effectively access the answerability.

## 3. Proposed Metric

This section presents the design and computation of our proposed evaluation metric PMAN, the Prompting-based Metric on ANswerability. The fundamental task of this metric is to determine if the reference answer is an appropriate answer to a generated question in accordance with the presupposed passage.

# 3.1. Prompt Design

Standard prompting describes the target task. Wei et al. (2022) discovered that providing a Chain-of-Thought (CoT) outlining the steps to perform a reasoning task would significantly improve LLMs' performance. Thus, we investigate the effectiveness of CoT for our evaluation task. For our case, we provide step-by-step instructions asking a GPT-based LLM to: (1) answer a question itself; (2) compare its answer to the reference answer; and (3) give the final assessment by responding with "YES" or "NO." Figure 1 presents the framework of our CoT prompting, consisting of three sections. When we adopt CoT, we use all sections. When not, we exclude section C from the prompt.

| Test sample set       | Assessing model | CoT          | Ans<br>Precision | werable<br>Recall | )<br>F1 | Non-a<br>Precision | nswera<br>Recall | ble<br>F1 | Accuracy | # of valid responses |
|-----------------------|-----------------|--------------|------------------|-------------------|---------|--------------------|------------------|-----------|----------|----------------------|
|                       | GPT-4           | <b>√</b>     | 1.000            | 0.880             | 0.936   | 0.893              | 1.000            | 0.943     | 0.940    | 100                  |
|                       | GPT-4           |              | 1.000            | 0.820             | 0.901   | 0.847              | 1.000            | 0.917     | 0.910    | 100                  |
| Manual (non-"yes/no") | GPT-3.5         | $\checkmark$ | 0.976            | 0.800             | 0.879   | 0.831              | 0.980            | 0.899     | 0.890    | 100                  |
| Manual (non- yes/no ) | GPT-3.5         |              | 1.000            | 0.520             | 0.684   | 0.676              | 1.000            | 0.806     | 0.760    | 100                  |
|                       | Llama2          | $\checkmark$ | 0.880            | 0.880             | 0.880   | 0.880              | 0.880            | 0.88      | 0.880    | 100                  |
|                       | Llama2          |              | 0.980            | 0.980             | 0.980   | 0.980              | 0.980            | 0.980     | 0.980    | 100                  |
|                       | GPT-4           | ✓            | 0.902            | 0.920             | 0.911   | 0.918              | 0.900            | 0.909     | 0.910    | 100                  |
|                       | GPT-4           |              | 0.978            | 0.880             | 0.926   | 0.891              | 0.980            | 0.933     | 0.930    | 100                  |
| Manual ("yes/no")     | GPT-3.5         | $\checkmark$ | 0.623            | 0.760             | 0.685   | 0.692              | 0.540            | 0.607     | 0.650    | 100                  |
| Manual ( yes/110 )    | GPT-3.5         |              | 1.000            | 0.020             | 0.039   | 0.505              | 1.000            | 0.671     | 0.510    | 100                  |
|                       | Llama2*         | $\checkmark$ | 0.870            | 0.833             | 0.851   | 0.790              | 0.833            | 0.811     | 0.833    | 48                   |
|                       | Llama2          |              | 0.806            | 0.580             | 0.674   | 0.672              | 0.860            | 0.754     | 0.720    | 100                  |
| Model (non-"yes/no")  | GPT-4           | <b>√</b>     | 0.810            | 0.979             | 0.887   | 0.976              | 0.788            | 0.872     | 0.880    | 100                  |
| Model ("yes/no")      | GPT-4           | ✓            | 0.851            | 0.966             | 0.905   | 0.939              | 0.756            | 0.838     | 0.880    | 100                  |

Table 1: Reliability evaluation of PMAN (Llama2\* was evaluated only on 48 valid responses). Manually created samples are balanced: half of the questions are answerable, while model-generated samples unbalanced data: 48 answerable vs. 52 non-answerable for non-"yes/no", 59 vs. 41 for "yes/no".

## 3.2. Metric Computation

We initially set the decoding temperature parameter to 0. For each question, we instruct an LLM to give an assessment. The assessment is valid if it contains "YES" or "NO." If not, we increase the decoding temperature and request the model to regenerate a response until it validly assesses the question. Consequently, each question is likely associated with a valid assessment containing "YES" or "NO." Then, we calculate the PMAN score as the percentage of questions with "YES" assessments among all assessed questions.

# 4. Reliability of the Metric

We experimentally evaluated the reliability of PMAN with manually created and model-generated samples.

# 4.1. Manually Created Test Samples

We manually created the test samples using the HotpotQA (Yang et al., 2018) dataset. A sample consists of a passage and a gold question-answer pair on it. We exactly used the randomly selected samples in the dataset for answerable questions. We used forged samples for non-answerable questions, in which the original question and answer are replaced with those from another randomly chosen sample.

Although "yes/no"-type questions cover only six percent of HotpotQA (Yang et al., 2018), in preliminary investigations, we observed that LLMs tend to perform quite differently on this type of question. Therefore, we created 100 samples for both "yes/no" and non-"yes/no" types and conducted separate tests with them. We evaluated manually created samples using the following three different

LLMs, i.e., GPT-4 (gpt-4-0613), GPT-3.5 (gpt-3.5-turbo-0613), and Llama2 (llama-2-70b-chat), with two variations of *with* and *without* CoT for each.

According to the results shown in the upper half of Table 1, (1) PMAN can demonstrate an acceptable performance, especially with a reliability of more than 90% when using GPT-4; (2) CoT prompting is effective for GPT-3.5 but not so much for the other two models; and (3) PMAN can be used with different LLMs. However, in the case of Llama2, nearly 60 percent of responses to "yes/no" type questions were invalid, i.e., could not provide an expected response, even after multiple re-generations.

#### 4.2. Model Generated Test Samples

The high capability of assessing manually created high-quality samples cannot necessarily guarantee alignment with humans for model-generated questions. Low-quality questions could fool a metric.

Thus, we further tested PMAN with 100 questions generated by EQG (Su et al., 2020), SQG (Pan et al., 2020), CQG (Fei et al., 2022), and GPT-4 with the prompt shown in Figure 2. We generated 100 questions for "yes/no" and non-"yes/no" types using pairs of passages and corresponding reference answers randomly selected from the HotpotQA dataset. The answerability labels of generated questions were determined by the authors. For the non-"yes/no" type, each model provided 25 questions. For the "yes/no" type, SQG provided 50 instead of EQG, as it did not handle the "yes/no" type. The results in the lower half of Table 1 demonstrate a strong alignment between PMAN using GPT-4 with CoT and human evaluations.

Generate a question from the passage delimited by triple backticks that can be answered by the answer delimited by triple dashes, where answering the question requires reasoning over multiple sentences in the passage.

Passage: " 'passage contexts" '
Answer: --question contexts--

Question:

Figure 2: Prompt used in GPT-based question generation

|     | •    |      |      |      |      |      | PMAN                 | НМ |
|-----|------|------|------|------|------|------|----------------------|----|
| EQG | 40.2 | 26.7 | 19.7 | 15.2 | 35.3 | 20.5 | 42                   | 40 |
| SQG | 40.7 | 27.3 | 20.1 | 15.4 | 36.9 | 20.3 | 41                   | 33 |
| CQG | 50.8 | 37.8 | 30.2 | 24.8 | 46.0 | 25.1 | 76                   | 69 |
| GQG | 38.8 | 25.4 | 18.5 | 14.3 | 35.1 | 21.9 | 42<br>41<br>76<br>97 | 97 |

Table 2: Evaluation results. BL, RG, and MTR stand for BLUE, ROUGE, and METEOR, respectively. HM is human assessment of answerability.

| Passage | Heinrich August Marschner (16 August 1795 - 14 December 1861) was the most important composer of German opera between Weber and Wagner. Carl Maria Friedrich Ernst von Weber (18 or 19 November 1786 - 5 June 1826) was a German composer, conductor, pianist, guitarist and critic, and was one of the first significant composers of the Romantic school. |
|---------|---|
| Ref. Q. | Heinrich Marschner was a composer who performed in the time frame after one of the first significant composers in what school of work? (Answer: Romantic)   |
| EQG     | heinrich august marschner was the most important composer of german opera between weber and a german composer who was one of the first significant composers of what school?  |
| SQG     | what genre did both heinrich marschner and karl von weber share ?   |
| CQG     | heinrich marschner and carl maria von weber both were composers of what school?   |
| GQG     | What school of music was Carl Maria Friedrich Ernst von Weber, a significant German composer who came before Heinrich August Marschner, associated with?  |

Table 3: Examples of generated questions.

# 5. Application of the Metric

We applied PMAN to the MQG models' output along with other metrics. An MQG model we implemented with GPT-4 achieved SOTA in generating answerable questions.

### 5.1. Dataset

We used the HotpotQA dataset (Yang et al., 2018). The size of the test set is 6,972. Answering each question in the dataset necessitates the ability to

reason over supporting sentences merged in a passage from two distinct Wikipedia documents.

### 5.2. Models

The aforementioned three conventional models and a MQG model using GPT-4 were employed. EQG (Su et al., 2020) and SQG (Pan et al., 2020) were among the earliest work for the MQG task, while CQG (Fei et al., 2022) achieved the SOTA. We also implemented GQG, a MQG model using OpenAI GPT-4 (gpt-4-0613) with the prompt shown in Figure 2.

#### 5.3. Metrics

In addition to PMAN, three conventional metrics (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005)) were applied. PMAN was implemented on gpt-4-0613 using the prompt shown in Figure 1. All PMAN scores were calculated using the same 100 samples randomly selected from the test set. In contrast, the other metrics were calculated using the whole test set except for GQG, which was evaluated using the 100 samples mentioned above. The answerability of generated questions was also assessed by the authors.

### 5.4. Results

Table 2 shows the results. The rank agreement between PMAN (automatic assessment) and HM (human assessment) is perfect.

Concerning EQG, SQG, and GQG, there is a seemingly fair correlation between conventional metrics and PMAN. However, while BLUE and ROUGE evaluate GQG as the worst, GQG achieved the best answerability. We think this is because, as exemplified in Table 3, there is a significant qualitative leap between GQG and the other MQG models, which goes beyond the capability of the overlap-based metrics. Therefore, PMAN could be complementary to them or even replace them. Last but not least, as our implemented GQG model achieves the highest PMAN score, it would serve as a new strong baseline model for generating answerable questions.

<sup>&</sup>lt;sup>1</sup>These 100 samples are different from the 100 samples used in section 4

# 6. Conclusion

This paper highlighted an urgent issue in guestion generation: the absence of an effective automatic evaluation metric to assess whether the generated questions are answerable. We proposed a Prompting-based Metric on ANswerability (PMAN) to address this issue, leveraging recent LLMs. Experiments with manually created and model-generated samples demonstrated reliability and strong alignment with human evaluations. The experiments also suggested that the notable Chain-of-Thought (CoT) prompting might not be effective for the most advanced models in some tasks. A comparison of our metric to other conventional metrics further indicated its potential to complement the conventional metrics and to guide future research in QG toward the generation of more answerable questions.

#### 7. 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.
- Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.
- Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuan-Jing 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.
- Tianbo Ji, Chenyang Lyu, Zhichao Cao, and Peng Cheng. 2021. Multi-hop question generation using hierarchical encoding-decoding and context switch mechanism. *Entropy*, 23(11):1449.
- Tianbo Ji, Chenyang Lyu, Gareth Jones, Liting Zhou, and Yvette Graham. 2022. Qascore—an unsupervised unreferenced metric for the question generation evaluation. *Entropy*, 24(11):1514.
- 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.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Geval: NLG evaluation using GPT-4 with better human alignment. arXiv cs.CL 2303.16634.
- Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.
- Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt.
- Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4636–4647, Online. Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. arXiv cs.CL 2303.04048.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.
- Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring question-specific rewards for generating deep questions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2534–2546, Barcelona, Spain (Online). International Committee on Computational Linquistics.
- 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.
- Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada. 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: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.