Automatic Evaluation of Multiple-Choice Items for Reading Comprehension: Effects of Question and Distractor Categories

John S. Y. Lee[1], **Yin Poon**[1], **Shunjie Wang**[1], **Chu Kai Wah Samuel**[2]
[1]Department of Linguistics and Translation, City University of Hong Kong
[2]Graduate School, The Education University of Hong Kong
jsylee@cityu.edu.hk, poonyin@outlook.com, ShunjieWang@hotmail.com, chukaiwah@eduhk.hk

Abstract

Automatic generation of multiple-choice (MC) items for reading comprehension can support language learning by providing large amounts of practice materials. To enable rapid development of MC generation models, automatic assessment is essential since it is time-consuming to manually evaluate question and distractor quality. Although Text Informativity (TI) has been adopted as an automatic evaluation metric, the ability of Large Language Models (LLMs) to estimate the TI scores of different categories of questions and distractors has not yet been thoroughly analyzed. This paper investigates LLM performance in calculating TI scores for the range of questions and distractors defined in the PIRLS (Progress in International Reading Literacy Study) and STARC (Structured Annotations for Reading Comprehension) frameworks. We show that automatically estimated TI scores may result in systematic preferences for some question and distractor categories, and recommend that TI scores be used for within-category comparisons only.

Keywords: multiple-choice items, text informativity, question generation, distractor generation

1. Introduction

A multiple-choice (MC) item for reading comprehension consists of a passage, a question, and a number of *answer options*, which must include one *key* (i.e., the correct answer) and several *distractors*. Automatic generation of MC items not only reduces teachers' workload (Cheung et al., 2023), but also provides language learners with large amounts of exercises and practice materials. Advances in artificial intelligence have led to generation models that can produce high-quality MC items for various text genres (Elkins et al., 2024; Kalpakchi and Boye, 2023; Xiao et al., 2023; Wang et al., 2022). However, evaluation methodology has mostly relied on manual assessment of item quality, which is extremely time consuming.

To facilitate rapid development of MC generation models, Text Informativity (TI) has been proposed as an automatic evaluation metric (Säuberli and Clematide, 2024). Using a Large Language Model (LLM) for TI calculation assumes the model's competence in reading comprehension, but this assumption may not hold in the face of sophisticated questions and distractors, such as those recommended in reading comprehension assessment research (King et al., 2004; Mullis and Martin, 2019).

This paper investigates LLM performance in estimating TI scores for the range of questions and distractors defined in the PIRLS (Progress in International Reading Literacy Study) (Mullis and Martin, 2019) and STARC (Structured Annotations for Reading Comprehension) (Berzak et al., 2020) frameworks. We show that the estimated TI scores may lead to systematic preferences for some question and distractor categories. To mitigate this bias, it is recommended that TI scores be used for within-category comparisons only.

2. Question dataset

According to the International Association for the Evaluation of Educational Achievement, reading comprehension questions should address four comprehension processes, as defined in the PIRLS standards (Mullis and Martin, 2019):

Retrieval The answer is explicitly given in a text span in the passage.

Inferencing Answering the question requires inferences about ideas or information that is not explicitly stated in the text.

Integrating Answering the question requires comprehension of the entire passage, or at least significant portions of it.

Evaluation The answer involves judgement about some aspect of the text, and is not necessarily found in the passage.

The three latter categories, which require deeper understanding of the text, are known as *higher-order* questions.

While large-scale MC datasets are available for evaluating LLM performance (Hendrycks et al., 2021; Li et al., 2024), they do not focus on reading comprehension and are not annotated with question or distractor categories. We harvested a set of 390 questions for reading comprehension in Chinese that have been manually annotated with

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 170–174
11 May 2026.

PIRLS		STARC	
Question category	# questions	Distractor category	# distractors
Retrieval	103	In-span	486
Inferencing	147	Out-of-span	486
Integrating	69	Out-of-text	486
Evaluation	71		

Table 1: Breakdown of PIRLS question categories (Section 2) and STARC distractor categories (Section 3) in our dataset of human-crafted MC items

PIRLS categories.[1] They consist of 100 questions from public examinations and 290 questions taken from online PIRLS exercises.[2]

3. Distractor dataset

Educators have advocated the use of distractors that reveal the test-taker's misunderstanding of the passage, which can provide more informative assessment (King et al., 2004). To this end, the STARC framework proposes a taxonomy of distractors that reflect increasing degrees of miscomprehension (Berzak et al., 2020):[3]

Category B (In-span) is a distractor that is based on a misinterpretation of the *critical span*, i.e., the text span in the passage that is relevant to the question.

Category C (Out-of-span) is a distractor derived from a text span that is outside the critical span and is irrelevant to the question.

Category D (Out-of-text) is a distractor based on external knowledge or common sense, without textual support in the passage.

We used the OneStopQA dataset (Berzak et al., 2020), which contains 163 short passages and 3 questions per passage. For each of the 486 questions, besides the key, one In-span distractor, one Out-of-span distractor and one Out-of-text distractor are provided.

4. Evaluation metric

Text Informativity (TI) estimates the degree to which an MC item measures the test-taker's comprehension of the given passage (Säuberli and Clematide, 2024). Two scores need to be computed:

Answerability Given a passage, a question and an answer option, whether the human judge (or the LLM in automatic evaluation) can correctly label the answer option as `true`/`false`. Answerability should be *high* for well-designed MC items with clear questions about the passage and unambiguous answer options.

Guessability Same as above, except that the judge is not given the passage. Guessability should be *low* for well-designed MC items, since it should be impossible to determine whether an answer option is `true`/`false` without the passage. Questions that are not tied to the passage, and distractors that are obviously false, would lead to higher guessability.

The TI score is defined as the difference between the answerability score and the guessability score. Hence, a high TI score means the MC item has answer options whose correctness can be determined (high answerability), but only based on the content of the passage (low guessability).

5. Research Questions

Automatic evaluation is critical for rapid development of MC generation algorithms. According to a study on a German dataset (Säuberli and Clematide, 2024), GPT-4 was able to assign higher TI scores to human-crafted MC items than machine-generated items. However, the study was agnostic to the nature of the underlying questions and distractors. Research in reading comprehension assessment has emphasized the importance of using questions that require a variety of skills (Section 2) and distractors that reveal a range of comprehension levels (Section 3). It is therefore important to ascertain whether LLMs can reliably calculate TI scores for all categories of questions and distractors. To address this issue, this paper seeks to answer the following research questions:

RQ1 *How well can LLMs estimate the TI score (answerability and guessability) of MC items with different question and distractor categories?* An LLM that is not competent with higher-order questions (Section 2), for example, would underestimate their answerability.

RQ2 *Does TI systematically favor any question category or distractor category?* Some distractors are designed to appeal to common sense rather than to the content of the passage (Section 3). They may have lower guessability since they would likely appear to be true when considered without the passage.

[1] Our question dataset can be downloaded from https://github.com/pypoon/PIRLS_ZH_Dataset

[2] These exercises were accessed from: https://read.smes.tyc.edu.tw/smes/PIRLS/ ; https://www.cacler.hku.hk/hk/content/basic/5675 ; https://drive.google.com/file/d/1QTTaarqMJRvy7wU_SzOmw RxE2raAMzgt/view

[3] The original scheme uses only the letters (B, C, D). The descriptive names are given to facilitate discussion.

Dataset	Answer-ability	Guess-ability	Text Informativity
PIRLS	0.9455	0.6468	0.2987
STARC	0.9534	0.5295	0.4239

Table 2: Automatically computed TI scores on human-crafted MC items

These two questions must be resolved in order to apply TI as an automatic metric for MC item evaluation. If TI penalizes challenging questions because of lower answerability (RQ1), or if it favors distractors with inherently lower guessability (RQ2), then it would not facilitate the selection of MC items with diverse question and distractor categories.

6. Experiments and Analysis

The Text Informativity (TI) scores of MC items were calculated with GPT-4o.[4] TI is based on the answerability score and guessability score. To determine the answerability and guessability of a triplet {<passage>, <question>, <answer>), the LLM was prompted to label the answer option as `true` or `false`. In our experiments, we used the prompt in Table 3 for answerability, and the prompt in Table 4 for guessability.

```
Text: <passage>
Question: <question>
Answer: <answer>

Based on the text above, is this answer correct (T) or incorrect (F)? Indicate only the letter T or F.
```

```
文本: <passage>
問題：<question>
答案：<answer>

根據上面的文本，這個答案是正確的(T)還是錯誤的(F)？僅輸出字母T或F。
```

Table 3: Prompt for answerability: the LLM is to label an answer option as `true`/`false` when given the passage and question in English (top) and in Chinese (bottom)

6.1. Overall results

Table 2 shows the overall answerability and guessability scores. Since all MC items in our datasets are manually crafted, the gold answerability should in principle be 100%. GPT-4o performed well on both the PIRLS (94.55%) and STARC (95.34%) datasets, achieving relatively high answerability.

When not given the passage, GPT-4o was less accurate in judging the answer options (64.68% for PIRLS and 52.95% for STARC). The substantial gap between the answerability and guessability scores suggests that GPT-4o did indeed understand the passages when tackling the MC items. While these results suggest that the LLM can approximate the ability of a human reader, the next sections will reveal variations in its competence for different question and distractor categories.

[4]Version 2024-05-13, via Azure OpenAI API. The temperature was set to 0.

```
The following question and answer are from a multiple-choice comprehension task about an unknown text.

Question: <question>
Answer: <answer>

Without knowing the text, only based on general knowledge, is this answer more likely to be correct (T) or incorrect (F)? Indicate only letter T or F.
```

```
以下問題和答案來自一篇未知文本的多項選擇閱讀理解題。

問題：<question>
答案：<answer>

在不知道文本的情況下，僅根據一般知識，這個答案更有可能是正確的(T)還是錯誤的(F)？僅輸出字母T或F。
```

Table 4: Prompt for guessability: the LLM is to label an answer option as `true`/`false` when given the question but not the passage, in English (top) and in Chinese (bottom)

PIRLS categories	Answer-ability	Guess-ability	Text Informativity
Retrieval	**0.9587**	0.6335	0.3252
Inferencing	0.9473	**0.5833**	**0.3639**
Integrating	0.9275	0.6812	0.2464
Evaluation	0.9401	0.7641	0.1761

Table 5: Automatically calculated TI scores on human-crafted MC items with various PIRLS question types (Section 6.2)

6.2. Effect of Question Categories

Table 5 shows the automatically calculated TI scores for questions belonging to each PIRLS category, using the dataset described in Section 2.

STARC categories	Answer-ability	Guess-ability	Text Informativity
In-span	0.9527	0.4074	0.5453
Out-of-span	0.9877	0.4753	0.5123
Out-of-text	**0.9897**	**0.3786**	**0.6111**

Table 6: Automatically calculated TI scores for human-crafted MC items with various STARC distractor categories (Section 6.3)

Answerability. GPT-4o performed best on the Retrieval questions, accurately labeling 95.87% of the answer options as `true`/`false`. These questions should be easier than the higher-order questions, which require more advanced comprehension skills. In line with this expectation, the LLM's performance degraded to 94.73% for Inferencing, and further down to 92.75% for Integrating. Performance on the Evaluation questions was slightly higher (94.01%). These questions may require subjective judgment based on general knowledge (Section 2), to which GPT-4o is well exposed.

Guessability. For similar reasons, the LLM was most capable in judging the answer options for the Evaluation questions (76.41%) without reading the passage. Retrieval and Inferencing questions, which are most directly related to the content of the passage, had the lowest guessable scores.

Implications. To address RQ1, GPT-4o was most accurate in calculating the answerability of Retrieval and Inferencing questions. It underestimates the answerability of Integration questions because of their more challenging nature. Evaluation questions were similarly penalized because of their higher guessability.

To answer RQ2, we evaluated direct use of TI in question selection for MC items. For the 85 passages containing questions of all four PIRLS categories, we ranked the questions in each passage by TI score. Consistent with the observations above, there was a preference for the Retrieval and Inferencing categories: 56.5% of the highest-scoring questions were either Retrieval (23 passages) or Inferencing (25 passages). In contrast, there appeared to be a bias against the other two categories, with only 43.5% of the highest-scoring questions representing the Evaluation category (15 passages) or Integrating category (22 passages). This bias can hinder the construction of a well-balanced assessment item set with a variety of question categories (Mullis and Martin, 2019). To mitigate this issue, the TI score should be used for selecting questions only within the same PIRLS category, so that score differences could be attributed solely to question quality and not to question category.

6.3. Effects of Distractor Categories

Table 6 shows the automatically calculated TI scores attained by distractors in each STARC category, using the dataset described in Section 3.

Answerability. While GPT-4o succeeded in labeling most distractors as `false`, its performance varied according to distractor categories. For a competent reader, an Out-of-text distractor should be the least plausible, since it is not supported by the content in the passage; in contrast, an In-span distractor, which is based on the critical span, should be the most plausible. Dovetailing with this expectation, GPT-4o was most capable of recognizing Out-of-text distractors as `false` (98.97%), followed by Out-of-span distractors (98.77%). It was most often misled by In-span distractors (95.27%).

Guessability. Out-of-text distractors are aimed at low-proficiency students who, unable to understand the passage, judge the answer option based on general knowledge rather than the passage. Consistent with this design, when not allowed to read the passage, GPT-4o most often failed to label Out-of-text distractors as `false`, leading to low guessability (37.86%). It had greater success in judging Out-of-span distractors as `false` (47.53%) since they are derived from irrelevant text spans and often contain implausible content.

Implications. To address RQ1, on the one hand, GPT-4o was least accurate in calculating answerability for In-span distractors, which are designed to be plausible even for a competent reader. On the other hand, it may favor Out-of-text distractors by assigning them lower guessability (labeling them as `true`), since they are designed to appeal to students who judge them based on common sense.

To answer RQ2, we investigated the degree to which TI may penalize In-span distractors and favor Out-of-text ones using a TI-based selection criterion. This criterion requires a distractor to be both answerable, i.e., labeled by the LLM as `false` when the passage is available; and not guessable, i.e., labeled by the LLM as `true` when the passage is not available. Among the distractors in our STARC dataset, 61.11% of the Out-of-text distractors met this criterion, compared to only 55.56% of the In-span distractors and 51.44% of the Out-of-span distractors. This shows that naive application of answerability and guessability could lead to an overuse of Out-of-text distractors in MC items, at the expense of the other distractor categories. These results reinforce our recommendation to use TI scores only for within-category comparisons.

7. Conclusions

We have presented an in-depth analysis on automatic evaluation of MC items using Text Informativity (Säuberli and Clematide, 2024) on English

and Chinese datasets annotated in the PIRLS and STARC frameworks. Results showed that GPT-4o underestimates the answerability of question and distractor categories that are more challenging, and that some categories have inherently lower guessability. Naive use of TI can lead to systematic preferences for more straightforward questions and Out-of-text distractors. It is recommended that practitioners use TI scores to compare only questions and distractors that are within the same category.

8. Bibliographical References

B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, D. Kulkarni, C. S. Seow, R. Wong, and M. T.-H. Co. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*, 18(8):Article e0290691.

Sabina Elkins, Ekaterina Kochmar, Jackie Chi Kit Cheung, and Iulian Vlad Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proc. 14th Symposium on Educational Advances in Artificial Intelligence*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proc. ICLR*.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proc. 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, page 477–491.

K.V. King, D. A. Gardner, S. Zucker, and M. A. Jorgensen. 2004. *The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics*. Pearson.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, page 11260–11285.

Ina V. S. Mullis and Michael O. Martin. 2019. *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement.

Andreas Säuberli and Simon Clematide. 2024. Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models. In *3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, page 22–37.

Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk. 2022. Towards Human-Like Educational Question Generation with Large Language Models. *Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science*, 13355.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page 610–625.

9. Language Resource References

Yevgeni Berzak and Jonathan Malmaud and Roger Levy. 2020. *STARC: Structured Annotations for Reading Comprehension*. Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL).