Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Automatic Distractor Generation in Multiple-Choice Questions Using Large Language Models with Expert-Informed Distractor Strategies

Yusei NAGAI[a*], Masaki UTO[a*]
[a]*The University of Electro-Communications, Japan*
*{nagai, uto}@ai.lab.uec.ac.jp

Abstract: In recent years, automatic generation of reading-comprehension questions with artificial intelligence has attracted considerable attention. In particular, producing high-quality distractors remains a critical challenge when generating multiple-choice questions (MCQs). Recent studies have increasingly employed large language models (LLMs) to generate distractors for MCQs. However, prior research has relied solely on the implicit, black-box knowledge of LLMs and has seldom exploited human expertise in distractor design. Therefore, in this study, we propose an LLM-based distractor-generation method that explicitly incorporates expert-informed distractor strategies, which represent typical heuristics used by human experts when crafting distractors. Experiments demonstrate that our method produces distractors of higher quality than those generated by previous approaches.

Keywords: Large language models, question generation, distractor generation, deep learning, natural language processing

1. Introduction

An effective way to develop reading comprehension skills is to provide learners with diverse reading passages accompanied by comprehension questions tailored to each passage. However, manually creating large numbers of such questions for various reading passages is time-consuming and labor-intensive. Recent studies have therefore explored the use of artificial intelligence to automatically generate reading comprehension questions (Chan et al., 2022; Tomikawa et al., 2024). Automatic question-generation methods can be used to create a variety of question types, among which the generation of multiple-choice questions (MCQs) is particularly prominent (Dutulescu et al., 2024; Gao et al., 2019; Maity et al., 2024; Shuai et al., 2023; Yu et al., 2024).

A crucial challenge in automatic MCQ generation is the generation of high-quality distractors. There are several criteria that high-quality distractors must generally satisfy. For example, (1) the distractors should not be easily identified and eliminated as incorrect options, and (2) the distractors should not be semantically equivalent or overly similar to the correct answer (Dutulescu et al., 2024). Therefore, creating distractors that satisfy these criteria is crucial for generating effective MCQs.

Numerous studies have been conducted with the aim of improving distractor quality in MCQs (Dutulescu et al., 2024; Gao et al., 2019; Maity et al., 2024; Shuai et al., 2023; Yu et al., 2024). For example, one approach generates candidate distractors using large language models (LLMs) and subsequently filters or ranks these candidates to remove inappropriate options (Dutulescu et al., 2024). Another approach involves multi-step prompting, in which distractors and questions are generated sequentially through structured prompts (Maity et al., 2024). However, these existing methods rely solely on the implicit, black-box knowledge embedded within LLMs without explicitly utilizing expert knowledge employed by human experts when creating distractors. Consequently, distractors produced

by these automatic methods may differ from those that human experts typically create, potentially reducing their appropriateness or effectiveness within specific contexts.

Therefore, in this study, we propose a novel distractor-generation method that explicitly incorporates human expert knowledge. Our approach first systematically organizes expert-informed distractor strategies, which reflect the common intentions and heuristics that human experts use when creating distractors. Next, our method uses a fine-tuned neural model to select the most suitable strategies for a given reading passage, question, and correct answer. Finally, an LLM guided by the selected strategies generates contextually appropriate distractors. Experiments show that our method yields higher-quality distractors than previous approaches.

2. Related Works

Dutulescu et al. (2024) proposed a method for automatically generating MCQs using LLMs that combines knowledge bases such as WordNet and DBpedia with the T5 model (Text-to-Text Transfer Transformer). Their approach generates candidate distractors, filters out inappropriate ones, and ranks the remaining candidates to select effective distractors. Yu et al. (2024) introduced a retrieval-augmented generation framework that incorporates external knowledge sources. Their method leverages the semantic relationships between words retrieved from knowledge bases to enhance the relevance and quality of generated distractors. Furthermore, Maity et al. (2024) proposed a multi-stage prompting method based on the chain-of-thought paradigm. In their approach, an LLM is prompted in four sequential stages: (1) paraphrasing the passage, (2) extracting keywords, (3) generating a question based on those keywords, and (4) generating distractors.

Although these methods have demonstrated improvements in distractor generation, they rely primarily on the implicit, black-box knowledge in LLMs. As noted in the Introduction, they do not explicitly incorporate expert-informed distractor strategies, which limits their ability to consistently produce distractors that align with those created by human experts.

3. Proposed Method

To overcome this limitation, the present study proposes an LLM-based distractor-generation method that explicitly leverages expert-informed distractor strategies. The proposed method comprises four steps: 1) classification of expert-informed distractor strategies, 2) construction of a strategy-labeled question dataset, 3) training of a strategy selection model, and 4) distractor generation. The first three steps constitute the preparation phase, while the final step corresponds to the actual distractor generation. Each of these steps is described in detail in the following subsections.

3.1 Classification of Expert-Informed Distractor Strategies

The initial preparation step involves systematically identifying and categorizing expert-informed distractor strategies by thoroughly reviewing prior research (Freedle and Kostin, 1991; Goodrich, 1977; King et al., 2004; Terao, 2019) that provided explicit guidelines or methodologies for distractor creation in MCQ design. Specifically, we carefully extracted, grouped, and synthesized these guidelines into distinct and coherent strategy categories. The resulting comprehensive set of strategies is presented in Table 1.

3.2 Construction of a Strategy-Labeled Question Dataset

The second preparation step involves annotating each distractor in an existing reading comprehension MCQ dataset with a corresponding strategy label based on the classification

Table 1. Expert-informed distractor strategies

	Strategy	Description
1	Use of Opposite Facts	Generating distractors that convey a meaning directly opposite to the correct answer.
2	Use of Irrelevant Facts	Generating distractors based on information unrelated to the topic or context of the reading passage and the correct answer.
3	Incorrect Combination of Facts	Generating distractors by incorrectly combining separate facts from different parts of the reading passage to create plausible yet incorrect options.

of expert-informed distractor strategies. The annotation process is conducted using Llama3, an open-source LLM, as described below.

Let an MCQ dataset be denoted as $\mathcal{D} = \{(c_n, q_n, a_n, \boldsymbol{d}_n) \mid n \in \mathcal{N}\}$, where c_n, q_n, a_n, and $\boldsymbol{d}_n$ represent a reading passage, a question, a correct answer, and a set of K distractors, respectively, for the n-th MCQ. The distractor set is defined as $\boldsymbol{d}_n = \{d_{nk} \mid k \in \mathcal{K}\}$ with $\mathcal{K} = \{1, \ldots, K\}$, and $\mathcal{N} = \{1, \ldots, N\}$ denotes the index set for MCQs with a size of N. To handle each distractor individually, we convert $\mathcal{D}$ into $\mathcal{D}' = \{(c_n, q_n, a_n, d_{nk}) \mid n \in \mathcal{N},\ k \in \mathcal{K}\}$. For each record in $\mathcal{D}'$, Llama 3 assigns the appropriate expert-informed distractor strategy, yielding the labeled dataset $\mathcal{D}_\mathrm{t} = \{(c_n, q_n, a_n, d_{nk}, y_{nk}) \mid n \in \mathcal{N},\ k \in \mathcal{K}\}$, where y_{nk} denotes the label of the expert-informed distractor strategy from Table 1 assigned to d_{nk}.

3.3 Construction of a Strategy Selection Model

The third preparation step is to construct a strategy-selection model that determines which strategies should be employed for a specific triple consisting of reading passage, question, and correct answer. We implement it as a classifier based on BERT (Bidirectional Encoder Representations from Transformers) The classifier receives the concatenated text of the reading passage, question, and correct answer as input and then outputs the label of the appropriate distractor-generation strategy. The classifier is trained on the dataset $\mathcal{D}_\mathrm{t}$.

3.4 Distractor Generation Leveraging Expert-Informed Strategies

After the above preparation steps, the proposed method generates distractors for an arbitrary reading passage c, question q, and correct answer a, using an LLM. Specifically, the strategy-selection model first predicts the most suitable distractor-generation strategy for the input triplet (c, q, a). Then, given the selected strategy y and the same triplet, a prompt is constructed using the template shown in Table 2, and this prompt is fed into Llama3 to generate distractors.

The prompt incorporates few-shot examples pertinent to the selected strategy, retrieved from $\mathcal{D}_\mathrm{t}$ by measuring their similarity to the input triplet $\{c, q, a\}$. The procedure for selecting similar examples is as follows:

1. The given triple (c, q, a) is input into SimCSE-BERT (Simple Contrastive Learning of Sentence Embeddings BERT) to obtain an embedding vector. Embedding vectors for all samples in $\mathcal{D}_\mathrm{t}$ are also precomputed in the same manner.
2. The cosine similarity between the embedding vector of the input triplet and those of the samples in $\mathcal{D}_\mathrm{t}$ labeled with the selected strategy y is computed. The three most similar samples $\{(c_i^{(\mathrm{e})}, q_i^{(\mathrm{e})}, a_i^{(\mathrm{e})}, d_i^{(\mathrm{e})}) \mid i \in \{1, 2, 3\}\}$ are chosen as few-shot examples.

4. Experiments

This section describes the experiments conducted to evaluate the performance of the proposed method. Our experiments used the RACE dataset (Lai et al., 2017), a well-known dataset for reading-comprehension MCQs. Each question comprises a passage, a question,

Table 2. Prompt template for generating distractors

You are an expert in creating multiple-choice questions for reading comprehension. You are provided with a set consisting of reading passage, a question, and a correct answer. The question and correct answer are both related to the content of the reading passage. Your task is to create a distractor (an incorrect option) based on the reading passage, question, and correct answer according to the following strategy. {*The description of expert-informed distractor strategy* y} During generation, please follow the steps below: 1. Understand the above strategy thoroughly. 2. Carefully read and understand the provided reading passage, question, and correct answer. 3. Generate a distractor following the strategy, considering the given reading passage, question, and correct answer. 4. Output ONLY the distractor you generated. The input passage, question, and correct answer are given below: ### Reading passage: c, Question: q, Answer: a
The following are the 3-shot examples (for $i \in \{1, 2, 3\}$): ### Reading passage: $c_i^{(e)}$, Question: $q_i^{(e)}$, Answer: $a_i^{(e)}$, Distractor: $d_i^{(e)}$

the correct answer, and three distractors. The experimental procedure was as follows:

1. We split the RACE training dataset into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$ in a 9:1 ratio, then constructed $\mathcal{D}_{\text{t}}$ from $\mathcal{D}_{\text{train}}$ following the procedure in Section 3.2.
2. We trained the strategy selection model following the procedure in Section 3.3.
3. For each of 100 randomly selected instances from $\mathcal{D}_{\text{test}}$, the proposed method generated one distractor according to the procedure in Section 3.4.
4. The generated distractors were evaluated manually on the following two criteria.
   - **Incorrectness:** This criterion indicates whether the option is actually incorrect. We assigned a score of 1 when the option was indeed incorrect; otherwise, the score was 0.
   - **Plausibility:** This criterion indicates whether the option cannot be dismissed without reading the passage. We assigned a score of 1 when recognizing the option as incorrect required proper comprehension of the passage; otherwise, the score was 0.
5. For comparison, we conducted the same experiments on the following two methods:
   - **Baseline:** A method that generates distractors without providing either expert-informed distractor strategies or few-shot examples
   - **Proposed Variant:** A variant of the proposed method in which distractors are generated using randomly selected few-shot examples, while the optimal strategy is still selected by the strategy selection model.

Table 3 presents the mean and standard deviation (in parentheses) of each score across the 100 generated distractors. The table shows that the proposed method obtained the highest values for both evaluation criteria. To determine whether these differences were statistically significant, we compared the methods using the Kruskal–Wallis test. The test revealed a significant difference in plausibility at the 1% significance level. Subsequent Dunn post-hoc tests with Holm correction confirmed significant plausibility differences (1%) between the proposed method and both the baseline and the proposed variant. These results indicate that the proposed method with appropriate few-shot examples produces higher-quality distractors that serve as effective and misleading options.

Table 3. Experimental results

Criteria	Proposed	Baseline	Proposed Variant
Incorrectness	**0.97** (0.17)	0.93 (0.25)	0.95 (0.21)
Plausibility	**0.56** (0.49)	0.20 (0.40)	0.32 (0.46)

5. Analysis

For a more detailed analysis, we assessed the generated distractors using item response theory (IRT), a family of probabilistic models widely used in educational and psychometric measurement to analyze the relationship between an examinee's latent ability and their item responses. IRT enables the estimation of item parameters, such as difficulty and discrimination, as well as examinee abilities on a common scale, offering more precise measurement than classical test theory. Based on this IRT framework, along with a virtual-examinee approach (Tomikawa et al., 2024; Uto et al., 2024), we conducted the following evaluation to assess the generated distractors:

1. We first built 59 virtual examinees as question-answering (QA) systems, following Tomikawa et al. (2024). Specifically, they were constructed using various pretrained neural models trained on the RACE validation dataset with different sample sizes, so that their accuracies spanned a wide range. We then collected their correct and incorrect responses on MCQs in $\mathcal{D}_{\text{train}}$, and estimated their abilities with the Rasch model, a widely used model in IRT.
2. For each record in $\mathcal{D}_{\text{test}}$, we generated three distractors: one using the baseline method, one using the proposed method, and one using the proposed variant method. Treating each set of three distractors together with the corresponding reading passage, question text, and correct answer as a single question, we collected responses from the 59 virtual examinees for each.
3. We converted the responses to distractor-level data, treating every distractor as an individual item and each response as an indicator of whether the examinee selected it. Using these data, we fitted the two-parameter logistic model, another widely-used IRT model, by fixing examinee abilities to their Rasch estimates, and obtained the discrimination parameter for each distractor.

Figure 1 plots the average absolute discrimination against the correct ratio threshold: the x-axis gives the maximum correct ratio (e.g., $x = 0.5$ covers questions with a correct-answer rate < 0.5), and the y-axis shows the mean absolute discrimination for the distractors of the corresponding questions. Note that we reported the absolute discrimination values, whereas all discrimination estimates were originally negative values because higher-ability examinees were less likely to choose distractors. The three curves correspond to the distractors produced by the baseline, the proposed method, and the proposed variant, respectively.

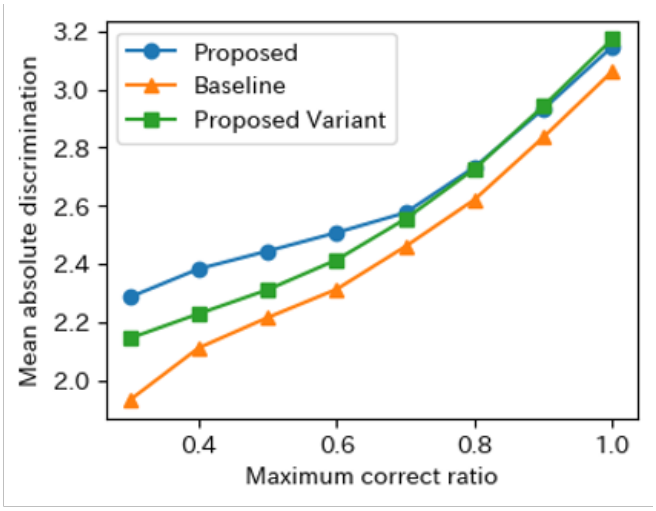

Figure 1. Discriminatory power of the generated distractors

The results show that distractors generated by the proposed method achieve the highest absolute discrimination, especially on difficult questions (i.e., those with low correct-answer rates). For easier questions, where the correct-answer rate is high, the gap between methods

narrows because distractors play a smaller role. Overall, these findings confirm that the proposed method produces distractors that better differentiate examinees' abilities, enabling automatically generated MCQs to diagnose reading-comprehension proficiency more accurately and to yield finer-grained ability estimates.

6. Conclusion

In this study, we proposed a method for generating high-quality distractors for multiple-choice reading-comprehension questions. Human evaluation and IRT analysis demonstrated that our approach yields distractors with higher quality and greater discriminatory power. Future work will proceed along three directions. First, because our experiments were limited in scale and detail, we will evaluate the method on more diverse datasets in greater detail. Second, the current taxonomy of distractor strategies remains coarse; therefore, we plan to refine it into finer subtypes and enable it to discover additional strategies automatically. Third, although the human evaluation currently relies on a binary scoring system, this may lack the granularity needed to capture nuanced differences in distractor quality. Future work will adopt a more fine-grained scoring rubric to help uncover deeper insights.

References

Chan, Y.-H., Chung, H.-L., & Fan, Y.-C. (2022). Keyword provision question generation for facilitating educational reading comprehension preparation. Proceedings of the 15th International Conference on Natural Language Generation, 196–202.

Dutulescu, A., Ruseti, S., Iorga, D., Dascalu, M., & McNamara, D. S. (2024). Beyond the obvious multi-choice options: Introducing a toolkit for distractor generation enhanced with NLI filtering. Proceedings of the Artificial Intelligence in Education, 242–250.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. ETS Research Report Series, 1–52.

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). Generating distractors for reading comprehension questions from real examinations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6423–6430.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. TESOL Quarterly, 69–78.

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

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale reading comprehension dataset from examinations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 785–794.

Maity, S., Deroy, A., & Sarkar, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. Proceedings of the 46th European Conference on Information Retrieval, 268–277.

Shuai, P., Li, L., Liu, S., & Shen, J. (2023). QDG: A unified model for automatic question-distractor pairs generation. Applied Intelligence, 53, 8275–8285.

Terao, T. (2019). Eigo bunsho dokkai koumoku ni okeru sakuranshi no sentakuritsu: Jukensha no tenkeiteki na goto ni chakumoku shite (distractor selection rates in English reading comprehension items: Focusing on examinees' typical errors) [Doctoral dissertation, Nagoya University] [(In Japanese)].

Tomikawa, Y., Suzuki, A., & Uto, M. (2024). Adaptive question–answer generation with difficulty control using item response theory and pretrained transformer models. IEEE Transactions on Learning Technologies, 17, 2186–2198.

Uto, M., Suzuki, A., & Tomikawa, Y. (2024). Question difficulty prediction based on virtual test-takers and item response theory. Workshop on Automated Evaluation of Learning and Assessment Content.

Yu, H. C., Shih, Y. A., Law, K. M., Hsieh, K., Cheng, Y. C., Ho, H. C., Lin, Z. A., Hsu, W.-C., & Fan, Y.-C. (2024). Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration. Proceedings of the Association for Computational Linguistics, 11019–11029.