A Non-Parametric Approach to Contextual Distractor Generation: Leveraging Error by Machine Reading Comprehension for Enhanced Distractor Generation

Abstract

In this work, we introduce a retrieval-augmented approach for generating distractors, leveraging hard negative passages from the open-domain question answering (ODQA) task. Unlike traditional methods reliant on dataset fine-tuning, our approach directly references teaching materials, enhancing practical application for educators crafting multiple-choice questions and allowing for seamless domain transitions without additional training. We adopt large language models (LLMs) for evaluation, tapping into their advanced comprehension capabilities to deliver a holistic assessment of distractor quality. Our experiments reveal superior performance over existing techniques, indicating a significant advance in the field.

1 Introduction

Effective distractors are crucial in automated assessment systems for evaluating the depth of a learner's understanding. Creating such distractors traditionally requires expert human judgment, making it a significant challenge in automated systems. Therefore, there has been an increasing focus on automating distractor generation through research (Liang et al., 2017, 2018; Chiang et al., 2022; Wang et al., 2023).

Our approach represents a departure from traditional distractor generation methodologies (Wang et al., 2023; Ren and Q. Zhu, 2021; Chung et al., 2020), which typically involves fine-tuning language models with specialized datasets. As illustrated in Figure 1, we introduce a non-parametric framework that utilizes the capabilities of Open-Domain Question Answering (ODQA) systems, specifically their proficiency in generating pseudo answers — distractors that are contextually aligned yet factually incorrect.

At the core of our method is the creative application of a Machine Reading Comprehension (MRC) Reader (Zhang et al., 2021), adapted in our framework as a *Contextual Distractor Synthesizer*. This Synthesizer is uniquely fed with pre-selected *hard negative passages* alongside the question stem. These hard negative passages are carefully curated text excerpts that closely align with the question stem in terms of topic, yet do not contain the correct answer, thereby serving as fertile ground for generating pseudo answers — plausible but incorrect distractors.

Distractor Synthesis Process: Upon receiving a hard negative passage and a question stem, the Contextual Distractor Synthesizer engages in generating pseudo answers. These are crafted to be contextually relevant to the question stem, leveraging the passage's content to ensure their plausibility as distractors. However, since these passages are specifically chosen for their lack of correct information, the resulting pseudo answers, while seemingly viable, are in fact incorrect, thus effectively functioning as high-quality distractors.

This integration of hard negative passages is a strategic choice in our methodology, balancing relevance with incorrectness to craft convincingly challenging distractors. Such an approach significantly augments the cognitive demands placed on learners, thereby elevating the overall quality of multiple-choice assessments.

The following features distinguish our DG method:

- **No Need for Fine-Tuning:** Setting our method apart from other DG models that require extensive fine-tuning, our approach operates without the need for specific training. This attribute allows for swift adaptability across various domains, as changing the knowledge corpus is sufficient to tailor distractor generation to different subject areas.

- **Control over Question Scope:** Our framework offers the unique ability to control the

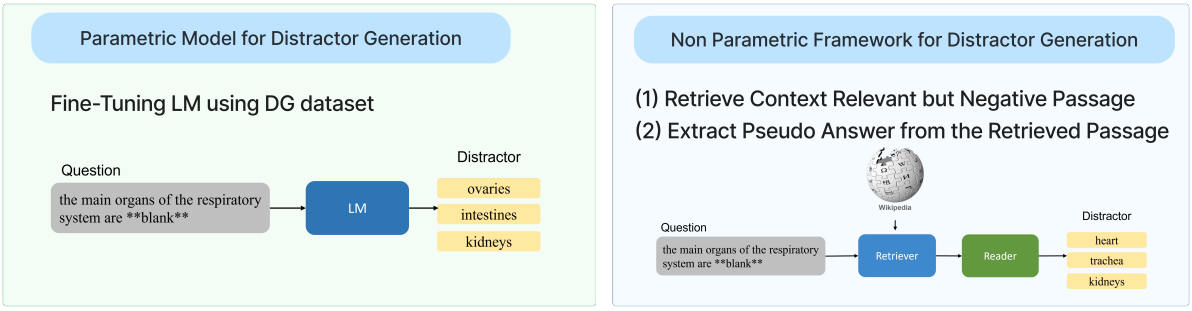

Figure 1: Contextual Distractor Synthesizer: This figure contrasts our proposed non-parametric framework (the left part of the figure) for distractor generation (right) with the existing fine-tuned Language Models (LM) approach (left) (Wang et al., 2023; Ren and Q. Zhu, 2021; Chung et al., 2020). Our framework operates by first retrieving contextually relevant passages that do not contain the correct answers. It then extracts pseudo answers from these passages with a machine reading comprehension model, utilizing them as potential distractor candidates.

thematic scope of the questions. By exchanging the data corpora, we can selectively focus on generating questions within specific topic boundaries. This aspect of our methodology resonates with the practical necessities of educational contexts, where educators often seek to limit question topics to relevant subject matter. This level of control over the question scope is a feature not readily achievable in conventional Text2Text models.

2 Related Work

The landscape of distractor generation (DG) research is currently delineated into two primary frameworks, each with its distinct methodologies and advancements.

Generating and Ranking (GR) Framework: This framework employs a general-purpose knowledge base to create a candidate set of distractors, followed by a feature-rich learning-to-rank model for distractor selection. The GR architecture operates in two stages: initial generation of candidate distractors and subsequent ranking based on semantic rules and linguistic features. There are two prevalent approaches within this framework: one utilizes a knowledge base (Ren and Zhu, 2021), and the other leverages a language model (LM) (Chiang et al., 2022). These approaches have marked a significant improvement over traditional rule-based methods (Liang et al., 2017, 2018), offering enhanced quality and diversity in distractors and are considered state-of-the-art in DG.

Text2Text Generation Architecture: The Text2Text generation framework formulates distractor generation as a Text2Text task, diverging from the GR approach. It involves concatenating the question stem with the correct answer and feeding this combined input into a generative language model (e.g., T5 or GPT). This architecture trains the model to generate a set of distractors in a direct and streamlined manner. Recent research adopting the Text2Text model (Wang et al., 2023) has set new benchmarks in distractor generation, demonstrating state-of-the-art performance.

Our Contribution - Contextual Distractor Synthesizer: Our work introduces a novel paradigm in the field of DG, distinct from both the GR and Text2Text architectures. At the core of our methodology is the use of a Machine Reading Comprehension (MRC) Reader, adapted as a *Contextual Distractor Synthesizer*. Unlike the GR framework, which relies on knowledge bases or LMs for initial distractor generation, our approach utilizes preselected hard negative passages that are contextually aligned but factually divergent from the correct answer. This method allows for the generation of pseudo answers that are inherently incorrect but contextually relevant, enhancing the cognitive challenge in assessments.

Moreover, our approach does not require the extensive training associated with the Text2Text models. By swapping out the knowledge corpus, our method easily adapts to various domains, offering flexibility and control over the scope of questions. This feature is particularly advantageous in educational settings where specificity and relevance are key.

3 Methodology

In this work, we introduce a distractor generation method that leverages incorrect answer patterns from Open-Domain Question Answering (ODQA)

to aid educators in crafting multiple-choice questions. Our approach utilizes 'hard negative' passages to inform a reader model, generating distractors without dataset-specific training. The process depends on a knowledge corpus, ensuring distractors are contextually aligned yet incorrect, to enhance the educational depth of assessments.

3.1 Problem Setting and Assumptions

The methodology assumes the following inputs:

- A knowledge corpus $\mathcal{C}$, consisting of text chunks relevant to the subject matter.
- A question stem Q, representing the query to be addressed.
- An answer A, the correct response to the question stem Q.

Additionally, the methodology utilizes two functions:

- A document retriever $\mathcal{F}()$, which fetches text chunks from the corpus $\mathcal{C}$.
- An extractive document reader $\mathcal{R}(Q)$, which extracts answers from a fetched text chunk given the question stem Q.

3.2 Algorithm Description

Please refer to the Appendix for details of our algorithm.

4 Experiment

4.1 Evaluation Metrics

We introduce GPT-4 Distractor Effectiveness Index (GDEI) to address the shortcomings of traditional token-based metrics in evaluating distractor quality. Token scores often overlook the semantic and contextual intricacies essential for effective distractors. In contrast, GDEI leverages GPT-4's advanced comprehension to holistically assess distractor sets, aligning with the nuanced demands of educational assessments.

4.2 GDEI Validation through Comparative Analysis

Our validation contrasts GDEI scores for educator-crafted ground truth distractors against those generated by T5 and BART models. The results favor ground truth distractors, showcasing GDEI's capability to discern quality reflecting human expertise in distractor design. This confirms GDEI's effectiveness as a nuanced evaluation tool for distractor generation.

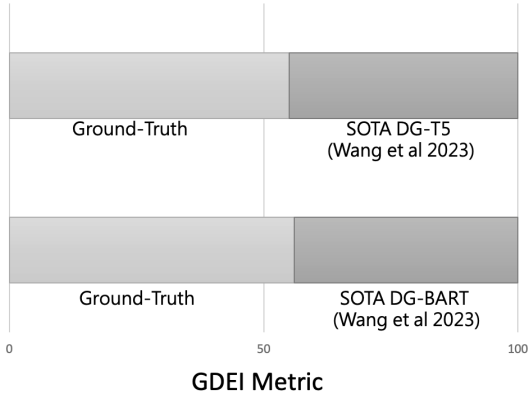

Figure 2: Enter Caption

4.3 Dataset

- **MCQ dataset**: MCQ dataset (Ren and Q. Zhu, 2021) is a cloze-style dataset, that includes the domains of science, vocabulary, common sense, and trivia. Each data is composed of a sentence containing **blank** of cloze stem, answer, and distractors.

- **MedmcQA dataset**: MedMCQA (Pal et al., 2022a) MedMCQA is a vast dataset of over 194k high-quality Multiple-Choice Questions and Answers for medical entrance exam preparation. It covers 2.4k healthcare topics and 21 medical subjects from AIIMS and NEET PG exams. The questions vary in length and complexity, and each sample includes a question, correct answer(s), additional options, and a detailed solution explanation.

In the in-domain experiment dataset, since the state-of-the-art (SOTA) methods utilized the MCQ dataset (Ren and Zhu, 2021) for training, we chose the MCQ test set as our testing dataset, comprising a total of 259 multiple-choice question sets. In the out-of-domain experiment dataset, we selected the medmcqa dataset, which is less related to the MCQ question type, to test the distractor generation ability of the SOTA model in out-of-domain scenarios. From the medmcqa (Pal et al., 2022b) test dataset, we selected 100 question sets with structures similar to those in the MCQ dataset.

4.4 Comparative Results Discussion

Our experimental analysis presents a comparative evaluation of our distractor generation method against state-of-the-art (SOTA) variants, namely T5 and BART, using two distinct datasets: MCQ and MEDMCQA. The GDEI metric serves as our evaluation criterion, quantifying the effectiveness of the generated distractors.

4.4.1 Performance on MCQ Dataset:

In the MCQ dataset, our method outperformed the SOTA models, achieving a GDEI score of 55 compared to 45 for SOTA DG-T5 and 47 for SOTA DG-BART. This indicates a clear preference for the distractors generated by our method, suggesting that our approach produces more contextually relevant and challenging distractors. The results underscore our method's capability to better mimic the characteristics of high-quality distractors, typically crafted by human experts.

4.4.2 Performance on MEDMCQA Dataset:

The MEDMCQA dataset further validates the superiority of our method. It attained remarkably high GDEI scores of 89 and 86, whereas the SOTA models scored significantly lower, with 11 for DG-T5 and 14 for DG-BART. The stark contrast in the scores on this dataset emphasizes the robustness of our method in a specialized domain. It also reflects the adaptability of our non-parametric framework, which, without the need for fine-tuning, can still generate highly relevant distractors even in niche areas of knowledge.

4.4.3 Implications of the Findings:

The consistent outperformance of our method across different datasets reinforces the importance of a non-parametric approach in the field of distractor generation. These findings highlight our method's ability to generate distractors that not only align more closely with the content and difficulty expected by educators but also adapt seamlessly across various domains of knowledge. The GDEI has proven to be a sensitive and reliable metric that captures the subtleties of distractor quality that may be missed by other evaluation methods.

In conclusion, the comparative analysis not only establishes the efficacy of our distractor generation method but also positions the GDEI as a valuable tool for future research and development in automated assessment systems. Our findings indicate a promising direction for enhancing the authenticity and educational value of multiple-choice questions through improved distractor generation.

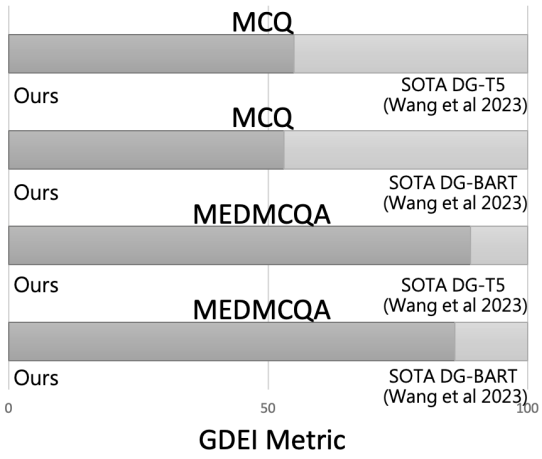

Figure 3: Enter Caption

Dataset	MCQ				MedicalQA			
	Train	Dev	Test	All	Train	Dev	Test	All
# of Questions	2088	233	259	2580	182822	4183	6150	193155

Table 1: The statistics of the training, development and test sets of MCQ and MedicalQA.

5 Conclusion

In conclusion, our DG method stands out for its unique features, notably the absence of fine-tuning requirements and the unprecedented control over question scope. Unlike other DG models, our approach operates efficiently without the need for extensive fine-tuning, allowing for swift adaptation across diverse domains by simply adjusting the knowledge corpus. The ability to selectively focus on generating questions within specific topic boundaries addresses the practical needs of educational contexts, providing educators with a level of control over question scope that conventional Text2Text models struggle to achieve. These distinctive characteristics position our DG method as a versatile and efficient tool for generating targeted and relevant questions across various subject areas.

6 Limitations

- **Assessment of Distractor Quality** Although the GDEI provides a more comprehensive evaluation of distractors, it still may not capture all aspects of distractor quality, such as the potential for a distractor to reinforce common misconceptions or to be pedagogically useful.

- **Dependence on Quality of Corpus:**The quality and diversity of the generated distractors are directly tied to the richness of the knowledge corpus used. If the corpus is outdated, biased, or lacks depth, the distractors may not be as effective or may inadvertently introduce inaccuracies.

References

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. Cdgp: Automatic cloze distractor generation based on pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A bert-based distractor generation scheme with multi-tasking and negative answer training strategies. *arXiv preprint arXiv:2010.05384*.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 284–290.

Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneaur, and C Lee Giles. 2017. Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In *Proceedings of the Knowledge Capture Conference*, pages 1–4.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022a. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022b. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4339–4347.

Siyu Ren and Kenny Q Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4339–4347.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor generation based on text2text language models with pseudo kullback-leibler divergence regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14506–14514.

A Algorithm Details

The distractor generation process employed in our methodology involves the following steps, each contributing to the effective creation of contextually relevant yet incorrect distractors for knowledge-based questions.

A.1 Retriever

The first stage of the process is the retrieval of relevant text chunks, a critical step in identifying suitable passages for distractor generation. We use the BM25 algorithm to retrieve relevant text chunks for distractor generation. We filter out passages that contain the correct answer and select 'hard negatives' that are closely related to the question but cannot provide an accurate answer. This ensures contextually relevant yet misleading content for effective distractor generation.

A.2 MRC Reader as Contextual Distractor Synthesizer

At this stage, the Machine Reading Comprehension (MRC) Reader is employed as a Contextual Distractor Synthesizer. This innovative approach deviates from traditional reading comprehension question-answering methodologies by shifting focus from extracting accurate answers to generating contextually relevant yet incorrect distractors, termed as *pseudo answers* as served as candidates for distractors. During the experiment, in the reader component of our framework, we select Llama2-7b and conduct training over two epochs using the SQuAD dataset as our reader.

A.3 Distractor Generation from Hard Negative Passages:

Initially, the MRC Reader scrutinizes hard negative passages, those related to the question but devoid of the correct answer. This careful selection of passages ensures that the generated pseudo answers are not only plausible but also challenging and contextually aligned with the question.

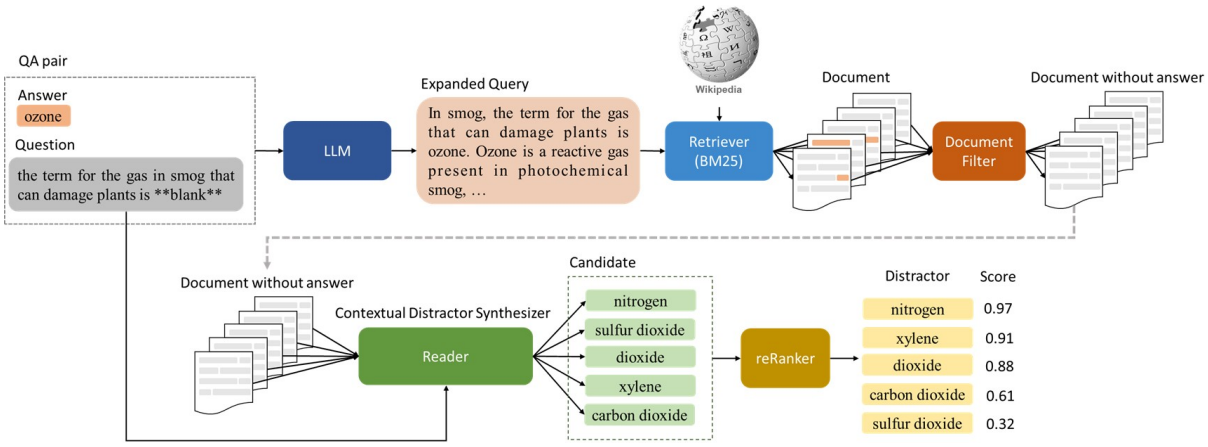

Figure 4: Overview of our approach

A.4 Formation of the Pseudo Answer Set (PAS):

PAS, or the Pseudo Answer Set, is a collection of all potential distractors generated by the MRC Reader. It is formulated based on two key parameters:

- k – the number of text chunks retrieved by the Reader, and
- h – the number of pseudo answers identified by the Reader for each text chunk.

Thus, PAS comprises a total of $k \times h$ pseudo answers, each representing a potential distractor derived from the hard negative passages.

A.5 Character-Level Rouge Score Evaluation:

To enhance the quality and diversity of PAS, each pseudo answer undergoes a character-level Rouge score evaluation. This evaluation assesses textual similarity and ensures that the pseudo answers are distinct from one another. Specifically, If a pseudo answer has a high Rouge score compared to any existing item in PAS, indicating a significant overlap, it is discarded to prevent redundancy.

A.6 Distractor Evaluator

The culminating phase of our distractor generation process is the evaluation of the pseudo answers by the "Distractor Evaluator." This component is critical for appraising the $k \times h$ pseudo answers produced by the Contextual Distractor Synthesizer, and its primary goal is to discern the most appropriate distractors for each question based on their relevance and plausibility.

Criteria for Evaluation: The evaluation is rooted in a series of criteria designed to ensure the quality and effectiveness of the distractors:

- **Semantic Closeness:** This criterion assesses each pseudo answer for its semantic similarity to the correct answer. The goal is to ensure that the distractors not only appear plausible but also present a substantial challenge to the test-takers.
- **Contextual Alignment:** The evaluator also examines how well each pseudo answer aligns with the context of the question. Effective distractors should be relevant to the subject matter of the question, beyond mere semantic similarity.

Features for Ranking Pseudo Answers: The input to the Distractor Evaluator comprises the question stem Q, the correct answer A, and the set of pseudo answers $PAS = \{PA_i\}$. The pseudo answers are ranked based on the following features:

- **Confidence Score s_0:** This score,

$$s_0 = p_{CDS}(PA_i|Q, C; \theta)$$

, reflects the confidence of the Contextual Distractor Synthesizer (CDS) in PA_i being a viable distractor.

- **Word Embedding Similarity s_1:** Calculated as

$$s_1 = 1 - \cos(\vec{A}, \vec{PA_i})$$

, this measures the cosine similarity between the word embeddings of A and PA_i.

- **Contextual-Sentence Embedding Similarity s_2:** This feature,

$$s_2 = 1 - \cos(\vec{Q}_{\otimes A}, \vec{Q}_{\otimes PA_i})$$

, evaluates the sentence-level similarity between the question stem filled with A and PA_i.

Final Selection of Distractors: Each pseudo answer PA_i receives a final score computed as a weighted sum of the individual scores with Min-Max normalization:

$$score(PA_i) = \sum_{\forall i} w_i \cdot \texttt{MinMax-Norm}(s_i)$$

The distractors that achieve the highest scores in this evaluation — the top-k scored items — are selected as the final distractors. This process ensures that the chosen distractors are both contextually relevant and sufficiently challenging.

Algorithm 1 Distractor Generation Algorithm

Input: Knowledge corpus $\mathcal{C}$ consisting of text chunks, question stem Q, answer A to Q
Assume:
    (1) a document retriever $\mathcal{F}()$ for fetching text chunks from $\mathcal{C}$
    (2) an extractive document reader $\mathcal{R}()$ for extracting an answer from a given text and a question
procedure GENERATEDISTRACTORS($\mathcal{C}, Q, A$)
    $\kappa \leftarrow \mathcal{F}(Q)$ ▷ Fetching Top-k text chunks
    for all $C_i \in \kappa$ **do**
        if A is in C_i **then**
            Remove C_i from κ
        end if
    end for
    Let PAS be Pseudo Answer Set = $\{\}$
    for all $C_i \in \kappa$ **do**
        $PseudoAnswer \leftarrow \mathcal{R}(C_i, Q)$
        if $PseudoAnswer$ not in PAS **then**
            Evaluate $PseudoAnswer$ with character-level Rouge score
            if $PseudoAnswer$ has low Rouge similarity with all items in PAS **then**
                $PAS.add(PseudoAnswer)$
            end if
        end if
    end for
    EVALUATEDISTRACTORS(PAS, Q, A)
    return Top-k scored items in PAS as final distractor set for Q and A
end procedure
function EVALUATEDISTRACTORS(PAS, Q, A)
    for all $PA_i \in PAS$ **do**
        Compute s_0, s_1, and s_2 for PA_i
        $score(PA_i) \leftarrow \sum_{j=0}^{2} w_j \cdot \texttt{MinMax-Norm}(s_j)$ ▷ Computing final score
    end for
    Sort PAS based on $score(PA_i)$ in descending order
    return Sorted PAS
end function

Please help me evaluate the results of the following distractor generation.
Question: {question}
Answer: {answer}
Distractor set 1: { distractor set 1 }
Distractor set 2: { distractor set 2}
From a teacher's perspective, which of the following sets do you think is better: Distractor set 1 or 2.
Please output according to the following format.
reason: _(up to 100 words)
choose:set 1 or set 2 or both are same

Figure 5: Prompt for GPT4 distractor evaluation.