ISSR: Iterative Selection with Self-Review Framework for Vocabulary Test Distractor Generation

Anonymous Author(s)

ABSTRACT

Vocabulary acquisition is essential to second language learning, as it underpins all core language skills. Accurate vocabulary assessment is particularly important in standardized exams, where test items evaluate learners' comprehension and contextual use of words. Previous research has explored methods for generating distractors to aid in the design of English vocabulary tests. However, current approaches often rely on lexical databases or predefined rules, and these methods frequently produce distractors that risk invalidating the question by introducing multiple correct options. In this study, we focus on English vocabulary questions from Taiwan's university entrance exams. We analyzed student response distributions to gain insights into the characteristics of these test items and provide a reference for future research. Additionally, we identified key limitations in how large language models (LLMs) support teachers in generating distractors for vocabulary test design. To address these challenges, we propose the iterative selection with self-review (ISSR) framework, which introduces an LLM-based self-review mechanism to ensure that the distractors remain valid while offering diverse options. Experimental results show that ISSR achieves promising performance in generating plausible distractors, and the self-review mechanism effectively filters out distractors that could invalidate the question.

KEYWORDS

English Vocabulary Test Design, Distractor Generation, Self-Review Mechanism, English Education Support, Large Language Models

1 INTRODUCTION

In second language learning, vocabulary acquisition plays a critical role as it serves as the foundation for all language skills, including listening, speaking, reading, and writing. Expanding one's vocabulary is a key factor in enhancing overall language comprehension [29]. A strong vocabulary enhances learners' ability to comprehend spoken and written texts, as a larger vocabulary enables more effective understanding. Additionally, understanding vocabulary is essential for mastering grammar, as knowing the meanings and functions of words aids in grasping sentence structures and applying grammatical rules. Therefore, vocabulary acquisition is fundamental to the overall development of language proficiency in second language learners.

Given the significance of vocabulary acquisition, it becomes crucial to develop effective ways of assessing learners' vocabulary knowledge. One common method is the use of vocabulary test items that require learners to select the correct word to fill a gap from a set of multiple choices. This type of assessment not only measures learners' ability to recognize the meaning of individual words but also evaluates their understanding of how words function in specific contexts. Furthermore, vocabulary questions have become an integral part of standardized English exams, especially in Asia, where they play a significant role in evaluating language proficiency. These exams include the General Scholastic Ability Test (GSAT),[1] Jitsuyo Eigo Gino Kentei (EIKEN),[2] Advanced Subjects Test (AST),[3] and the Test of English Proficiency developed by Seoul National University (TEPS).[4] These indicate the critical importance of well-designed vocabulary questions in assessing language proficiency. In this work, we aim to explore effective approaches for automatically designing English vocabulary test items that accurately assess learners' vocabulary comprehension. Automating test generation is expected to benefit both students and teachers. For students, it provides a greater variety of vocabulary contexts, helping them apply and understand words in different situations, which may contribute to improved retention. For teachers, it can reduce the burden of creating tests while maintaining consistent quality, allowing them to focus more on guiding student learning.

There are two common styles of vocabulary test questions. The first type, used in exams such as the GSAT, ELKEN, and TEPS, presents a sentence with one word omitted and offers four options, only one of which correctly fits the context. For example, consider the following sentence: "Posters of the local rock band were displayed in store windows to promote the sale of their concert tickets." in this case, the target word "concert" is masked, and the examinee must choose the correct word from a set of options, which includes distractors that do not fit the context. The second type, found in exams like the TOEFL iBT,[5] does not omit the target word. Instead, examinees are asked to select the option that most closely matches the meaning of the given target word. Since the second type provides a clear target for examinees to reference, the strategy for generating distractors may differ. For example, Susanti et al. [23] considered the target word itself when selecting appropriate distractors for the first type of exam. Despite the differences in format, both question types rely on the careful design of distractors to effectively assess vocabulary knowledge. In this paper, we focus on the first type of question, where generating distractors that test the examinee's ability to infer meaning from context is essential.

The quality of a vocabulary test question is based on designing a stem that effectively tests the examinee's understanding of a target word, followed by creating distractors that challenge the examinee to distinguish the target word from similar but incorrect options. By presenting plausible yet incorrect options, effective distractors challenge students to discern subtle differences in meaning, ensuring that they do not merely guess but instead demonstrate a thorough comprehension of the vocabulary and the context in which it is used. This meticulous design ensures a more accurate assessment of the student's language proficiency and understanding.

[1]https://www.ceec.edu.tw/en/xmdoc/cont?xsmsid=0J180519944235388511
[2]https://www.eiken.or.jp/
[3]https://www.ceec.edu.tw/en/xmdoc/cont?xsmsid=0J180520414679660023
[4]https://en.teps.or.kr/about_teps.html
[5]www.ets.org

Regarding the conditions required for a well-constructed vocabulary exam question, Heaton [11] identified several key elements for designing effective distractors: (1) The target word should have the same part of speech as the distractors. (2) The distractors and the target word should have a similar difficulty level. (3) The length of the target word and each distractor should be close. (4) Synonym pairs should be avoided in the options. However, manually designing distractors is labor-intensive. Automatically generating distractors based on a given stem and target word, offering suggestions for teachers to select from, has gained attention in recent research. Susanti et al. [23] propose a method to extract candidate distractors from lexical databases and predefined word lists (e.g., WordNet, JACET8000) and measure the semantic relatedness between the candidate distractors and the stem, serving as the basis for selection. However, this strategy for selecting distractors may be limited by the fixed range of words in the word lists, restricting the diversity and flexibility of the generated distractors. Liang et al. [17] propose a model trained on manually designed features, which allows for the generation of distractors that more closely resemble those found in real exam questions, thereby improving the quality and applicability of the distractors. Chiang et al. [7] construct a framework that uses neural network models to generate candidate distractors, followed by predefined rules to rank the best distractors, ensuring both the effectiveness and diversity of the generated options.

Although previous methods [7, 17] have achieved promising results, they have required a large amount of training data to develop effective models for generating distractors in English vocabulary tests. Obtaining such data is not always feasible, particularly for specialized vocabulary tests. This challenge presents the need to explore alternative approaches that do not rely on large-scale training data and are not limited by predefined dictionaries.

Recently, LLMs have demonstrated powerful semantic understanding and language generation abilities. Through in-context learning, combined with a few examples or specific prompting techniques like Chain-of-Thought prompting [18, 22, 26], these models can achieve excellent performance across various natural language processing tasks [4, 14], even without the need for additional training. Given these advantages, it is intuitive to consider leveraging LLMs to address the aforementioned challenges. However, it remains unclear whether LLMs can effectively perform the task of distractor generation. Therefore, this lead to the first research question (**RQ1**): Are LLMs capable of being directly utilized for generating distractors?

Using LLMs to generate distractors directly may pose several challenges. First, while LLMs are proficient in language generation, producing effective distractors requires precise control over their similarity to the correct answer. Distractors must be sufficiently misleading but not too close to the correct answer. Additionally, distractors should not align too closely with the stem, which could result in multiple correct answers. Furthermore, distractors need to be contextually appropriate and avoid any logical inconsistencies with the stem or other options. Another important consideration is the difficulty level of the distractors, as they must strike a balance between being misleading and not overly simple or difficult. This balance is crucial to ensure the test effectively differentiates learners of varying proficiency levels. Hence, we raise the second research question (**RQ2**): How can we ensure that the generated distractors maintain validity and do not introduce ambiguity or multiple correct answers?

To address the research question mentioned above, this paper investigates the English vocabulary test items in the GSAT exam, which is one of Taiwan's most significant university entrance exams. The GSAT is widely recognized for its ability to assess students' vocabulary proficiency, making it an ideal subject for our study. Its test items encompass a broad range of vocabulary and include distractors of varying difficulty levels, offering a comprehensive framework for evaluating vocabulary mastery. By focusing on key features such as the range of vocabulary, the relationship between distractors and correct answers, and common error patterns among students, we aim to provide a detailed analysis of the effectiveness and challenges posed by the existing test items. Given the direct impact of GSAT results on students' academic futures, our thorough examination of its vocabulary questions seeks to contribute not only to the refinement of vocabulary assessment methods but also to broader efforts in standardizing educational evaluation practices.

In this work, we propose the interative selection with self-review (ISSR) framework, which assists teachers in designing English vocabulary exams by generating and validating distractors. The framework consists of three modules: the candidate generator, distractor selector, and distractor validator. We leverage a pretrained language model (PLM) for generating contextually relevant distractors, and introduce an LLM-based "self-review" mechanism to ensure the question remains valid with only one correct answer. This framework is expected to reduce the manual effort required to design distractors while potentially enhancing the diversity and accuracy of test items, making them more reflective of students' vocabulary comprehension. Moreover, ISSR provides a flexible process, allowing for the integration of different advanced models into the framework. This adaptability enables the framework to evolve with advancements in language models, ensuring its versatility across various vocabulary exams and assessment needs. Additionally, ISSR does not rely on additional data for fine-tuning, further enhancing its efficiency and ease of use. With the ability to adjust prompts, this approach has the potential to adapt to a variety of vocabulary exams, making it a versatile tool for different assessment needs. The details of each module are discussed in the following section. In sum, our contributions in this paper are threefold:

- We investigate the challenge of using LLMs for automatic distractor generation in English vocabulary assessments, specifically focusing on generating contextually appropriate distractors that maintain validity and avoid ambiguity.
- In exploring the limitations of LLMs in this task, we identified key challenges and developed the ISSR framework based on these findings.
- Experimental results show that the distractors generated by ISSR perform well, and the proposed self-review mechanism effectively filters out invalid distractors.

2 RELATED WORK

Recent research on the automatic generation of distractors typically divides the process into two steps [7, 21, 23]: (1) Generating a set of distractor candidates, and (2) Ranking these candidates to select the most plausable distractors. Various methods have been applied to each of these steps. In the following, we will review related work by discussing these two steps separately.

Distractor Candidates Generation. This step aims to generating a broad set of distractor candidates, which will be further filtered in the next step. Susanti et al. [23] propose a method for obtaining and ranking distractors from lexical databases and dictionaries, specifically for the second type of exam described in Section 1. Their approach generates distractors based on the target word, stem, and correct answer. To maintain validity, distractors should resemble the correct answer in meaning but still be distinguishable. During the distractor candidates generation process, they retrieved candidates from the text passage by focusing on words with the same part of speech and tense, as well as sibling words from WordNet [9] and JACET8000 [12], a dictionary tailored for Japanese English learners. JACET8000 organizes 8,000 English words into eight levels of difficulty based on frequency in the British National Corpus,[6] supplemented by texts designed for Japanese students. They observed that most distractors in human-created vocabulary questions have a similar or nearly identical level of difficulty to the correct answer. Based on this findings, they collected words that closely match the difficulty level of the correct answer. If an insufficient number of distractor candidates were identified, the selection scope was expanded to include related terms of the target word from the Merriam-Webster Dictionary for consideration.[7]

Apart from vocabulary tests, there are open-domain multiple-choice questions that assess knowledge in specific areas such as science, common sense, and trivia. Ren and Zhu [21] focus on generating distractors for such questions, noting that many options in these datasets could be sourced from knowledge bases. They propose a method for selecting distractor candidates from WordNet and Probase [27]. One challenge they addressed was polysemy, where a word like "bank" could have multiple meanings. To resolve this problem, they apply context-dependent conceptualization [16], using a probabilistic topic model, LDA [3], to match the context with the most relevant concepts in the lexical database. They calculate the score of each distractor based on a probability distribution of concepts derived from both the target word and the question stem. The top N scoring distractors are then selected as the final candidates.

Previous work primarily utilizes lexical databases and dictionaries to select distractor candidates. In contrast, Chiang et al. [7] explore the use of pretrained language models (PLMs) for distractor generation, specially targeting on the CLOTH dataset [28], which consists of cloze-type questions created by teachers. Apart from the first exam type introduced in Section 1, the cloze-type questions present a paragraph with several words removed. Students are tasked with selecting the most appropriate word from four options to fill each blank. These questions assess not only vocabulary knowledge but may also include fill-in-the-blank tasks for prepositions and other grammatical elements. Given that PLMs are designed to predict words in masked positions or complete unfinished sentences, they fine-tune the models to generate plausible distractors for the question stem, rather than the correct answers.

[6]http://www.natcorp.ox.ac.uk/

[7]https://www.merriam-webster.com/

Distractor Candidate Scoring. In this step, candidates are ranked based on their suitability as effective distractors, with the goal of selecting the most competitive options. Susanti et al. [23] emphasize that effective distractors should be semantically similar to the target word, while remaining distinct from the correct answer to maintain question validity. They propose a ranking formula that considers semantic similarity and collocation, using GloVe word vectors [20] to calculate cosine similarity and NLTK to analyze collocation frequency between words.

Ren and Zhu [21] further advance the process by transforming distractor candidates into 33-dimensional feature vectors, which are then used to train ranking models, such as AdaBoost [10], LambdaMART [6], and other list-wise rankers. This approach allows for more sophisticated ranking and selection of distractors from a large pool of candidates, leveraging both point-wise and pair-wise ranking strategies.

Similarly, Liang et al. [17] propose neural network (NN)-based and feature-based models for distractor selection. They formulate the task as selecting the most plausible distractors from a given set. While utilizing the feature-based classifier model such as logistic regression, random forest [5] and LambdaMART [6], they employ an adversarial training framework (IR-GAN) [25] to enhance the NN model's performance. Their proposed cascaded learning framework, inspired by computer vision techniques [24], enables more efficient filtering and ranking of large distractor sets, demonstrating the effectiveness of feature-based models over NN-based models. Chiang et al. [7] propose a scoring mechanism that integrates PLM confidence scores with semantic similarity measures, including both contextual sentence embedding and word embedding similarity By assigning weights to different scoring methods, their framework identifies the top k distractors with the highest overall scores. This approach highlights the growing influence of pretrained language models in enhancing the accuracy and relevance of distractor generation. In summary, our work shares similarities with prior work in leveraging successful methods for distractor candidate generation and scoring. However, it diverges in its approach by incorporating LLMs to overcome limitations identified in previous studies. Specifically, we aim to address issues such as the over-reliance on dictionary-based distractor selection, the necessity for large training datasets, and the challenge of effectively filtering out distractor candidates that might also be plausible correct answers. By integrating LLMs, we seek to enhance both the flexibility and precision of distractor generation, while maintaining the strengths of earlier methods.

3 DATA ANALYSIS

In this study, we selected questions from the GSAT English exam as our dataset, with past exam questions available on the College Entrance Examination Center (CEEC) website.[8] The exam years we gathered range from 93 to 105, comprising a total of 195 questions. The GSAT English exam is designed to to assess whether candidates possess the fundamental academic abilities necessary for university education, serving as a preliminary screening criterion for university admissions and program selection. We consider these carefully curated exam questions to be a suitable benchmark for exploring

[8]https://www.ceec.edu.tw/xmfile?xmsid=0J052424829869345634

methods that support teachers in evaluating Taiwanese high school students' second language learning abilities. While the dataset is specific to Taiwan, the insights gained from this study may have broader applicability in similar educational contexts.

To develop a method for generating English vocabulary questions that closely align with those designed by teachers and to explore the characteristics of distractors that effectively mislead students, we analyze the features of English vocabulary questions in the following section. First, as vocabulary selection is a crucial aspect of vocabulary test, we will analyze the range of vocabulary chosen for the exam and how the difficulty of these words is evaluated. Next, we will examine the relationship between the target words and distractors in the dataset, with particular attention to the characteristics of distractors in questions with low pass rates. Finally, we will analyze the factors that influence students' selection of distractors, aiming to identify the key elements that diffuse students' choices.

3.1 Analysis from the CEEC Word-list

The CEEC has published a Senior High School English Word-list,[9] which records the vocabulary that Taiwanese high school students should understand before taking GSAT English test and AST English test. Each word in this list includes the word's difficulty level, and part of speech. The word list encompasses vocabulary categorized as nouns, verbs, adjectives, adverbs, articles, pronouns, conjunctions, and prepositions. Notably, adverbs formed by adding "-ly" to adjectives are not included separately. The vocabulary is classified into six levels of difficulty, with the classification based primarily on word frequency as recorded in the Cobuild English Dictionary [1] and the Cobuild English Dictionary for Advanced Learners [2]. While word frequency is the main factor for classification, additional rules are also applied to determine word difficulty levels.

Derivatives: Derivatives are typically considered more difficult than their root words, so in most cases the difficulty of derivatives is higher than that of the root. For example, the difficulty level of the word "true" is 1, while its derivative "truth" is level 2, and the word "truthful", which is a derivative of "truth", has a difficulty level of 3. However, if the frequency of derivatives is more common than the root, then the difficulty of the derivative may still be lower than that of the root. For example, "consideration" is regarded as higher frequency than it's root word "considerate", so the word "consideration" is leveled at three, while "considerate" is leveled at five.

Commonly Used Vocabulary in Culture: Since this exam is designed for Taiwanese students, a few low-frequency words in Cobuild English Dictionary from Chinese culture are also included. (e.g., "bamboo", "chopsticks", "dumpling")

To examine the difficulty distribution of vocabulary actually used in the GSAT English exam, we analyze the difficulty levels of the target words in the exam questions by checking the word difficulties of the target word in each question. Since some target words in the exam may appear in different forms (e.g., due to tense changes), we lemmatized the target words using the spaCy[10] toolkit

[9]https://www.ceec.edu.tw/SourceUse/ce37/ce37.htm
[10]https://spacy.io/

Table 1: Examples of Vocabulary in Various Difficulty Level

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
alone	beef	bandage	ashamed	considerate	anonymous
bathroom	arrange	crown	aluminum	astronaut	contradiction
already	calendar	deposit	container	bruise	eloquence
eagle	fever	envy	emphasis	contemporary	humanitarian
always	interview	fur	inflation	immense	prosecution

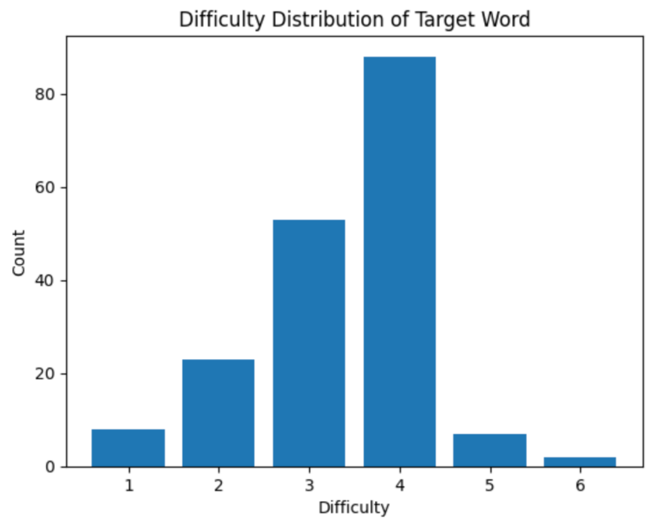

Figure 1: Difficulty Distribution of Target Word

before checking their difficulties to ensure that the base forms match the entries in the Senior High School English Word-list. Figure 1 shows the difficulty distribution of target words in the GSAT exam questions according to the Senior High School English Word-list. The x-axis represents the difficulty levels of the target words, while the y-axis indicates the number of words at each difficulty level. We find that the majority of words fall into difficulty levels 3 and 4, with approximately 60 and 80 words, respectively. Fewer words are distributed across difficulty levels 1, 2, 5, and 6. This indicates that the testing purpose of GSAT English exam is to test words in moderate difficulty, with only a few being particularly easy or difficult. Table 1 shows some example of words in level 1 to 6.

3.2 Relation between Distractors and Answer

Cosine similarity between distractors and answer: Cosine similarity is a commonly used metric for analyzing word similarity [7, 15, 23]. The higher the similarity between two words, the more likely they share the same semantic. Given two different words A and B, the cosine similarity is defined as follows:

$$\text{cosine\_similarity}(A, B) = \frac{\text{vect}(A) \cdot \text{vect}(B)}{\|\text{vect}(A)\| \|\text{vect}(B)\|}$$

In vocabulary tests, a distractor that is semantically similar to the target word is more likely to be a plausible option. However, if the similarity is too high, the question may become invalid, as both the target word and the distractor could be correct answers within the context. Hence, we measure the cosine similarity between the target word and the distractors to determine how similar a distractor should be to the target word in order to be considered a plausible choice. To compute the cosine similarity between distractors and the target word in this dataset, we use the `en_core_web_lg`[11] model from spaCy to convert the words into vectors and then compute their cosine similarity. The average cosine similarity of distractors in the GSAT English dataset is 0.290, with standard deviation equals to 0.131, indicating that the distractors exhibit a low to moderate level of similarity to the target word, without being too similar to invalidate the question.

To examine whether questions with low student pass rates exhibit low cosine similarity, we selected questions with pass rates below 60% and calculated the cosine similarity between the target word and its distractors. The results show that the average cosine similarity between the target word and the distractors in these questions is 0.257, with a standard deviation of 0.08. Our analysis indicates that in these high-error-rate questions, the cosine similarity between the target word and the distractors was not significantly higher. This suggests that high semantic relatedness between distractors and the target word is not a primary factor in misleading students. Moreover, determining an optimal level of cosine similarity for distractors remains inconclusive, as a distractor with high similarity to the target word may invalidate the question by providing multiple potential correct answers. Therefore, we opted not to incorporate cosine similarity between the target word and distractors into our framework at this stage.

Difficulty Difference Between Distractors and Answers: In Section 3.1, we analyzed the difficulties distribution of vocabularies select by GSAT as target word. In this section, we further investigate the difficulty levels of the words that teachers choose as distractors when assessing students' English vocabulary skills. As noted by Heaton [11], the relative difficulty of the distractors in relation to the correct answers is a crucial factor in determining the quality of the questions. If the distractors are significantly more difficult or easier than the correct answer, students are less likely to be misled, which makes it difficult to assess their true understanding of the target word's usage.

We analyze the difficulty difference between the target word and the distractors for each exam question. Similar to the method used in Section 3.1 for calculating the difficulty of target words, we first lemmatized the distractors using the `en_core_web_sm` model in the spaCy toolkit. We then determined their difficulty levels based on the Senior High School English Word List. After that, we compared the difficulty differences between the target words and their corresponding distractors. Figure 2 shows the difference between the distractors and the target word. The x-axis represents the difficulty levels of the target words, while the y-axis represents the number of words at each difficulty level. As shown in Figure 2, the difficulty difference between the target word and distractors in most questions falls within the range of 0 to 2 levels. Only a few questions exhibit a difficulty difference of 3 or 4 levels between the target word and the distractors.

We found that most questions maintain a difficulty level for the distractors that does not exceed one level higher than that of the answer. Since difficulty levels are classified based on word frequency, it is possible that words with similar frequency levels are more likely to be grouped together in the exam questions. To ensure that the generated distractors align with the nature of the exam questions we analyzed, our framework considers the difficulty difference between the distractor and the answer.

[11] https://spacy.io/models/en

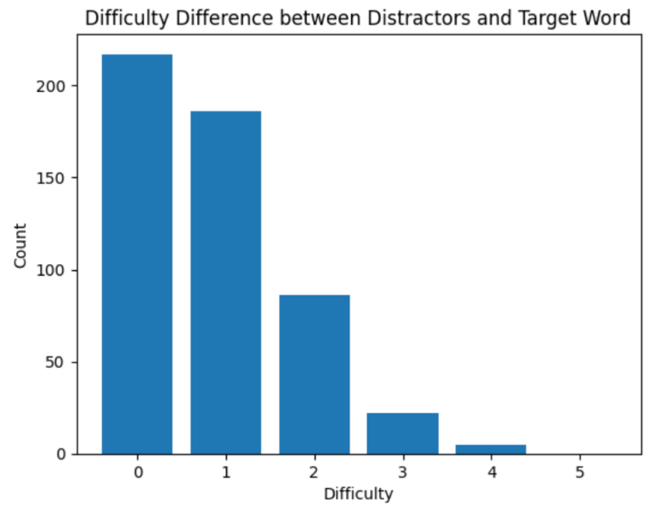

Figure 2: Difficulty Difference Between Distractors and Answers

3.3 Analysis of Question Types in Which Students Commonly Make Mistakes

Correlation Between Vocabulary Difficulty and Examinee Performance: The Senior High School English Word-list classifies words into different difficulty levels based on their frequency of occurrence. However, it remains uncertain whether relatively low-frequency words indeed present more challenges for students. Therefore, we collected the selection rates of each option by students as published by CEEC. For each vocabulary question, the CEEC provides data on the proportion of students who selected each option, referred to hereafter as the *selection rate*. The *pass rate*, in particular, denotes the selection rate of the correct option, representing the percentage of students who answered the question correctly. In this section, we analyze the relationship between the pass rates and the difficulty of the answers, where difficulty is defined according to the classifications in the Senior High School English Word-list.

Figure 3 shows the pass rates of students for questions across six difficulty levels. The Pearson correlation between question difficulty and student pass rate is -0.118, with a p-value of 0.113. This suggests that there is no significant correlation between overall pass rate and question difficulty. Based on our analysis and the exam statistics from the past few years, vocabulary words across all difficulty levels appear to be effective as test items for assessing students' mastery of English vocabulary.

The Role of Polysemous Words in Examinee Performance: One intuitive way that students may try to understand the meaning

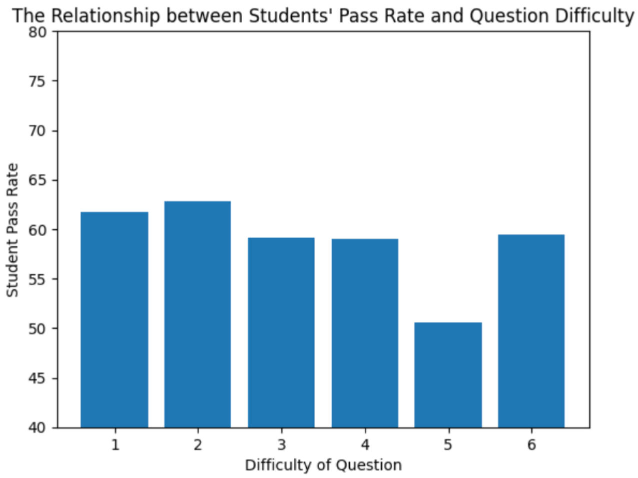

Figure 3: The Relationship between Students' Pass Rate and Question Difficulty

of English words is by using translation software to translate them into their native language. As a result, students may only become familiar with the most common meaning of the word, as translation software typically provides the most frequent or widely used meaning. This reliance on translation software can cause students to overlook other contexts in which the word is used, making it more difficult for them to recognize less common meanings when encountering the word in various contexts.

In this study, we aim to investigate whether it is particularly challenging for students when the correct answer in a vocabulary question is a polysemous word. Specifically, if a question tests a less common meaning of a polysemous word, are students more likely to choose the wrong answer due to unfamiliarity with that particular meaning? To explore this, we conducted an experiment to assess whether the Chinese meanings of English vocabulary words (i.e., correct answers to exam questions) align with those generated by translation tools.

First, we consulted WordNet to obtain all possible definitions of the target word. We then combined these definitions with the question stems and used `GPT-3.5-turbo-0613` to assess whether the meaning of the target word, as used in the question stem, aligned with any of its definitions. This process enabled us to identify which specific definition of the target word was being tested in the question. Next, we collected publicly available test explanations from textbooks, [12] which provide Chinese translations of the target words. These translations were written by teachers or English education experts who analyze the exam content and compile it into textbooks. Since teachers translate the target word into its appropriate Chinese meaning based on the context of the stem, we aim to assess whether the Chinese translation corresponds to the meaning of the English target word. Thus, we translated the Chinese meaning provided in the textbooks back into English. We then compared this back-translation word to the original target word to see if they corresponded. If the back-translated word from Google Translate differs significantly from the original target word, it may indicate that the target word is not commonly used to represent this specific Chinese meaning. For translations that did not match the target word, we further examined whether the meaning of the back-translated word aligned with the definition of the target word in this context. We asked `GPT-3.5-turbo-0613` to verify whether the back-translated word corresponded to the definitions initially obtained. If the back-translated word did not encompass the tested definition, it suggested that the target word was not commonly used to express this specific meaning in Chinese, resulting in a semantic mismatch.

The premise of this experiment is that students often rely on translation software to translate texts and learn English vocabulary. Therefore, we aimed to investigate whether the word generated by the translation software differs from the answer used in the exam. For questions where the back-translated word did not match the definition of answer, the average pass rate was 55.91%, whereas for questions where the back-translated word matched, the average pass rate was 60.01%. This suggests that students are more likely to answer incorrectly when a polysemous word is used in a less common sense, which may not be accurately captured by translation software. Specifically, in questions where the answer is used in an uncommon meaning, the pass rate was 4.10% lower than the overall average.

At this stage, we have not developed specific methods for designing questions that focus on polysemous words. Nonetheless, these preliminary findings suggest that further exploration is warranted regarding the role of polysemous words in exams and their impact on student performance, which we plan to address in future work.

4 ITERATIVE SELECTION WITH SELF-REVIEW FRAMEWORK

In this work, we propose the ISSR framework, which generates a set of distractors to assist teachers in designing English vocabulary tests. As shown in Figure 4, the framework is composed of three key modules: the candidate generator, the distractor selector, and the distractor validator.

4.1 Candidate Generator

Given a question stem and the target word, this module aims to generate a large number of candidate distractors for further selection. To ensure that the generated distractors possess the necessary qualities, such as being confusing yet plausible within the context, without being too easily dismissed, we adapted a pretrained language model (PLM) as our distractor generator. We chose to use a pretrained language model (PLM) for candidate generation rather than relying on lexical databases commonly used in previous work for several reasons. (1) A PLM is designed to generate words within a given sequence by understanding thecontext, producing outputs that are both coherent and semantically meaningful. This makes PLMs highly effective as word generators, as they can produce words that are contextually relevant and thus plausible as distractor candidates. (2) Unlike predefined rules that extract candidates from lexical databases like WordNet and Probase, PLMs can effectively generate a massive quantity of distractor candidates.

[12]https://elearning.sanmin.com.tw/englishsite/download/download.htm

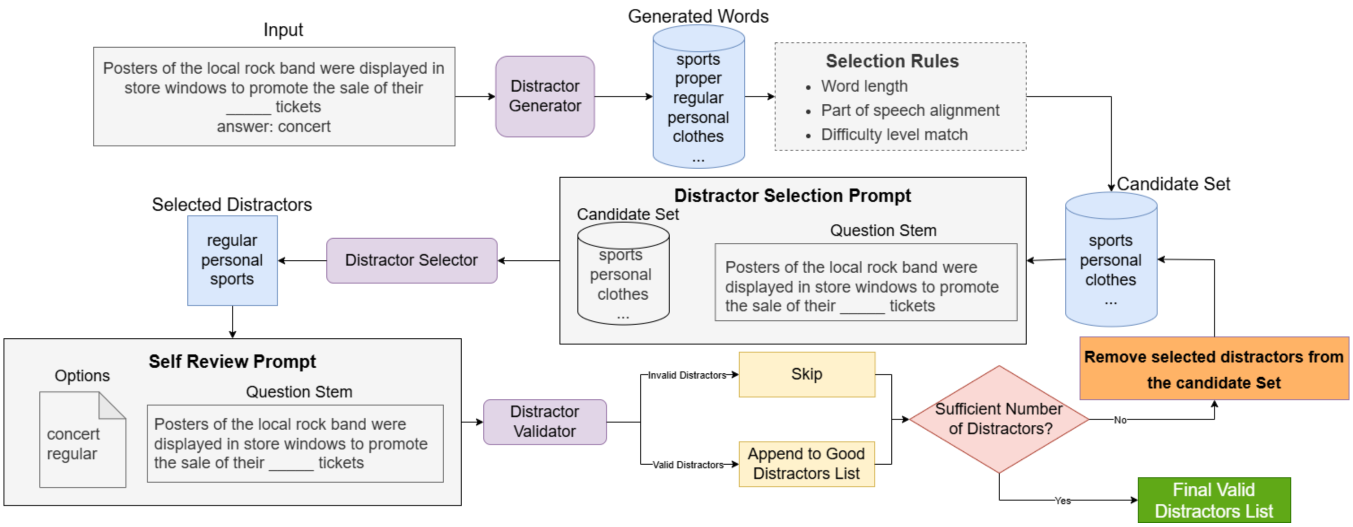

Figure 4: The ISSR Framework

We leverage the BERT-base model named CDGP-CSG, trained on the CLOTH dataset proposed by Chiang et al. [7]. Although CDGP-CSG was initially designed for generating distractors in a different context, it has the potential to generate distractors that appear plausible within the context of vocabulary exams. The target word in the given question stem is masked and input into the BERT model for fill-in-the-blank processing. Subsequently, the generated results are subjected to additional conditional filtering to ensure their suitability. Based on the analysis presented in Section 3 and the criteria for effective distractors identified by Heaton [11], the filtering rules are as follows:

- The length difference between the answer and the distractor should not exceed 2 characters.
- The answer and the distractor should share the same part of speech.
- The difficulty levels between the answer and distractor should be closely matched.

To verify whether the answer and the distractors share the same part of speech, we insert the generated distractors back into the stem and utilize NLTK toolkit to analyze their parts of speech. Additionally, to ensure that the difficulty levels of the answer and distractors are comparable, we reference the vocabulary list provided by the Taiwan CEEC. After lemmatizing both the answer and the distractors, we retain only those distractors with a difficulty difference of no more than 1. This process is designed to extract up to 50 distractors for the candidate pool.

4.2 Distractor Selector

Due to the ineffectiveness of directly generating distractors by the LLM, we opt instead to present a pool of candidate distractors and have the LLM select the most suitable ones. Further details on the limitations of direct distractor generation will be discussed in the following section. In this step, we input the generated candidate pool into the prompt and instruct the LLM to choose the top k most appropriate distractors. After obtaining several distractors generated by our candidate generator, a post-processing method is proposed to filter out those that are not suitable for the scope of the English vocabulary exam. Given a question stem, target word and a set of distractor candidates, we develop a distractor selector to select k appropriate distractors from distractor candidates. Previous research has primarily focused on using ranking models [17] and manually defined formulas [7, 23]. Although ranking models can identify appropriate for the questions, they not only require large amounts of training data but may also not be applicable across different exams scopes. The words ranked by these models may only be suitable for a specific difficulty level within a particular exam scope, making it difficult to apply them broadly across different exam requirements. Manually defined formulas may also not always optimal because it is difficult to define the ideal degree of semantic similarity between a distractor and the target word. A distractor with high semantic similarity to the target word may invalidate the question by creating multiple correct answers.

According to our analysis in Section 3.2 using GSAT English exam data, there is no significant correlation between the cosine similarity of distractors, indicating that semantic similarity may not be the primary factor in challenging examinees. To this end, we attempt to integrate an LLM into the framework for generating English vocabulary questions. By leveraging the LLM's powerful semantic understanding and natural language generation capabilities, we aim to utilize in-context learning to select appropriate distractors based on instructions without the need for additional training. An LLM-based distractor selector is incorperated into ISSR to select top k distractors from the candidate set.

4.3 Distractor Validator

The candidate distractors generated in the previous steps might also be plausible answers that could fill the blank in the question stem, which could result in distractors becoming potential correct answers. However, since the design of the test requires only one

Table 2: Performance of Distractor Generation

Method	F1@3	F1@10	NDCG@3	NDCG@10	NDCG@30
CDGP	0.51%	1.10%	1.19%	3.15%	5.95%
GPT-3.5 w/ Zero-Shot	0%	0.32%	0%	0.59%	1.34%
GPT-3.5 w/ Few-Shot	0.35%	0.56%	0.85%	1.71%	3.27%
Direct Selection	0.35%	0.24%	0.68%	0.85%	1.40%
ISSR	**1.55%**	**2.07%**	**3.57%**	6.31%	**9.82%**
w/o Self-Review	1.04%	1.91%	3.11%	**6.78%**	7.44%

correct answer among the four options, further filtering is necessary to ensure that these distractors cannot function as valid correct answers. To ensure that the selected distractors are not viable answers for the stem, we propose a self-review mechanism to assess their validity. Specifically, we formulate new questions by pairing each distractor with the correct answer, creating a binary choice. This allows us to evaluate the suitability of each distractor by presenting the options to the LLM. If the model selects the distractor as the correct answer, it suggests that the distractor is unsuitable for the question, as it would lead to the problem becoming invalid due to the presence of more than one correct answer. We also explore other self-review methods, and the experimental results will be discussed in detail in the next section. Note that, if there are not enough qualified distractors (fewer than k), the selected valid and invalid distractors are removed from the candidate set, and the distractor selector is invoked to select additional appropriate distractors from the remaining candidates.

5 EXPERIEMENTS

5.1 Experimental Setup

In the experiment, we utilized English vocabulary questions from the GSAT, spanning from 2006 to 2018. The dataset consists of 195 questions. We apply both zero-shot and few-shot prompting approaches in our experiments. For few-shot prompting, the first two questions were used as demonstrations for in-context learning, while the remaining 193 questions serve as the test set. In the zero-shot prompting setting, the same 193 questions were used directly as the test set without any prior demonstrations. To ensure we could extract the desired information from the LLM's responses, such as the distractors selected by the LLM, we structured the prompts with specific formatting instructions, guiding the LLM to follow a designated response format.[13] The temperature for all LLMs were set to 0.7. To verify that our ISSR framework improves the automatic distractor generation capabilities of LLMs, we compared the ISSR framework against the following baselines: CDGP, GPT-3.5 [19] and GPT-4o-mini [13]. We followed the original CDGP setup as specified, without any modifications. The GPT-3.5 model used in this comparison is `gpt-3.5-turbo-0125`, and GPT-4o-mini used in this comparison is `gpt-4o-mini-2024-07-18`.

5.2 Experimental Results

Table 2 presents the results of ISSR compared to other models. Since the primary goal of this work is to generate a sufficient number of high-quality distractors for teachers to select from, we extended the generated distractor count to 30. F-score and NDCG are adopted as the evaluation metrics. In addition to the baselines mentioned in Section 5.1, we assess the LLM's ability to select suitable distractors from the entire vocabulary, referred to as the "Direct Selection" method. Specifically, we instruct `gpt-4o-mini-2024-07-18` to select distractors directly from the vocabulary list provided by the Taiwan CEEC. The choice of `gpt-4o-mini-2024-07-18` as the selector is based on its reputation as a stronger LLM compared to `gpt-3.5-turbo-0125`, making it an ideal model for evaluating distractor selection due to its advanced language processing capabilities. If the LLM performs well under this approach, it would suggest that large language models have sufficient capability to handle wide-ranging distractor selection autonomously. As shown in Table 2, ISSR not only surpasses the direct use of `gpt-3.5-turbo-0125` for distractor generation and Direct Selection with `gpt-4o-mini-2024-07-18`, but also outperforms CDGP. This finding suggests that having a well-designed distractor candidate set plays a crucial role in generating high-quality distractors, emphasizing the importance of a well-curated candidate generation process.The findings indicate that ISSR is more effective in generating distractors that closely align with those found in actual exams, thereby offering educators a broader selection of viable distractor options. Furthermore, Table 2 shows that the integration of a self-review mechanism, which removes distractors that could potentially invalidate the question, contributed to an overall improvement in performance.

An interesting finding is that both zero-shot prompting and few-shot prompting using `gpt-3.5-turbo-0125` for distractor generation do not perform as well as CDGP on this task. This is likely due to the inherent instability of `gpt-3.5-turbo-0125` during generation. Our experiments show that when `gpt-3.5-turbo-0125` are tasked with generating a large set of distractors in a single round, they often produce repetitive content, as presented in Table 13. To tackle this issue, we experimented with generating a smaller number of distractors over multiple rounds, which partially reduced repetition. However, this approach requires modifying the prompt after each round to prevent the model from producing similar outputs, as `gpt-3.5-turbo-0125` tend to generate redundant content if the prompt remains unchanged.

In our implementation, we addressed this by adding explicit instructions to the prompt, directing the LLM to avoid repeating previously generated distractors, as demonstrated in Table 12. However, as restrictions on the reuse of prior outputs increase, `gpt-3.5-turbo-0125`'s generation can become increasingly erratic, resulting in repetitive or less relevant content. This limitation poses a significant challenge when a large number of unique distractors are required.

This finding indicates the advantages of the ISSR framework, which mitigates these issues by reframing the task for the LLM: instead of directly generating distractors, ISSR focuses on selecting suitable distractors from a pre-generated candidate set. This approach leverages the LLM's semantic capabilities to ensure that the selected distractors are both diverse and contextually appropriate. Further discussion and examples comparing ISSR to direct generation methods can be found in Appendix B.

[13]The prompts used in this work are presented in Appendix A.

Table 3: Performance of Using Different Candidate Generator

Candidate Generator	F1@3	F1@10	NDCG@3	NDCG@10	NDCG@30
None	0.35%	0.56%	0.78%	1.56%	2.46%
BERT-base-uncased	0.52%	1.12%	0.85%	2.69%	5.71%
CDGP-CSG	**1.55%**	**2.07%**	**3.57%**	**6.31%**	**9.82%**

Performance of using different candidate generator. The ISSR framework is highly flexible, allowing for the integration of various models as candidate generators, which are used to produce candidate distractors. In this experiment, we tested the impact of selecting different sources as the candidate generator on ISSR. The comparison involves three settings: (1) without a candidate generator, denoted as None, where distractors are directly generated using `gpt-3.5-turbo-0125` with zero-shot prompting, and a self-review process is applied to filter out invalid distractors; (2) a standard BERT-base model; and (3) a BERT-base model fine-tuned on the CLOTH dataset, referred to as the CDGP-CSG model. It is important to note that the ISSR framework incorporates predefined rules, as outlined in Section 4.1, to filter candidates generated by the candidate generator. For consistency and fairness, the same filtering mechanism has been applied uniformly across all candidate generators.

Table 3 presents the results of using different models as the candidate generators. The results show that the CDGP-CSG model achieves superior performance compared to the BERT-base model. In contrast, the standard BERT-based candidate generator exhibits weaker performance. Since BERT was trained on masked token prediction task and next sentence prediction task, it is intuitive that BERT will attempt to generate tokens that fit the given contexts, leading to the generation of suboptimal distractor candidates. The CDGP-CSG model, on the other hand, has fine-tuned BERT model to generate distractors on the CLOTH dataset. Although the CLOTH dataset differs from the GSAT in terms of exam scope, CDGP-CSG has learned to generate plausible distractors for a given question stem instead of merely generating words that fit the contexts.

Results of using Different LLMs for Distractor Selection. The pre-training datasets and architectures of different LLMs vary, leading to differences in their abilities to select distractors. In this experiment, we assess the capabilities of different LLMs in selecting distractors. The comparison involves three different models with varying parameter sizes: (1) `gpt-3.5-turbo-0125`, (2) `Llama3-8B` [8], and (3) `Llama3-70B`. For each model, both few-shot and zero-shot settings are evaluated and the CDGP-CSG model is used as the candidate generator.

Table 4 shows the results of using different LLMs to select distractors. The results demonstrate that `gpt-3.5-turbo-0125` with zero-shot prompting outperforms both `Llama3-8B` and `Llama3-70B`, despite the significant difference in their parameter sizes. Notably, `Llama3-8B` exhibits performance comparable to `Llama3-70B`, suggesting that model size does not necessarily correlate with improved performance in this task. Additionally, Table 4 indicates that the distinction between few-shot and zero-shot settings has minimal effect on the performance of LLMs when selecting the most appropriate

Table 4: Result of using different LLM for distractor selection

Model	F1@3	F1@10	NDCG@3	NDCG@10	NDCG@30
Llama3 8B w/ Zero-Shot	0.52%	1.36%	1.36%	3.60%	7.47%
Llama3 8B w/ Few-Shot	0.51%	1.12%	1.55%	3.32%	7.06%
Llama3 70B w/ Zero-Shot	0.69%	0.96%	1.62%	3.20%	7.68%
Llama3 70B w/ Few-Shot	0.86%	1.43%	1.95%	4.01%	8.49%
GPT-3.5 w/ Zero-Shot	**1.55%**	**2.07%**	**3.57%**	**6.31%**	**9.82%**
GPT-3.5 w/ Few-Shot	**1.55%**	1.59%	3.05%	4.99%	8.61%

distractors from a candidate set. The zero-shot setting slightly outperforms the few-shot setting, suggesting that the LLM is capable of effectively selecting distractors without requiring prior demonstrations. Based on these findings, we adopt `gpt-3.5-turbo-0125` with zero-shot prompting in the ISSR framework.

6 ANALYSIS AND DISCUSSION

6.1 Impact of Candidate Set Size

It may seem intuitive to think that providing more distractor candidates to the distractor selector will improve the outcome, as it gives the LLM more options to choose from. However, it remains unclear whether the LLM can accurately select proper distractors from a large candidate set. During the prompt design phase, we found that the size of the candidate set may influence the quality of the LLM's output, which ultimately affects the performance of the distractor selector. As the size of the candidate set increases, the likelihood of the LLM generating distractors which do not appear in distractor candidate set also increases.

To more precisely analyze the impact of candidate set size on the LLM's ability to effectively select distractors, we conducted an experiment to determine the optimal size for the candidate set. In this experiment, we utilized ISSR and varied the number of candidate distractors provided to distractor selector. Specifically, we extracted the stem and target word from the original teacher-designed test questions and generated multiple prompts, each corresponding to a different candidate set size generated by distractor generator. We then evaluated whether the distractor selector could accurately select distractors from within the provided candidate set. The LLM used in ISSR was `gpt-3.5-turbo-0125`.

Table 5 shows that the success rate of selecting three different distractors varies with different sizes of candidate sets. The success rate represents how accurately the distractor selector can choose distractors As shown in Table 5, we found a negative correlation between the size of the candidate set and the rate at which the LLM accurately selects distractors. While the LLM can reliably select distractors from a candidate set of size 50, its performance becomes more unstable as the candidate set size increases, ultimately resulting in a drop in accuracy to 90.67% when the candidate set size reaches 300. We hypothesize that the LLM's responses may not align with the original prompt's request to select suitable words from the distractors because a larger candidate set may disrupt the LLM's understanding that the current context still pertains to the candidate set. This can lead the LLM to infer possible relationships between distractors within the set, ultimately affecting the outcome. We conclude that reducing the candidate set size is necessary to

Table 5: Relationship Between Candidate Size and Successful Distractor Selection Rate

Candidate size	Selection Rate
300	90.67%
200	92.75%
100	97.93%
50	**98.79%**

Table 6: Performance of Various LLMs in Answering Vocabulary Questions

LLM	Accuracy
Llama3-70B	98.46%
Llama3-8B	95.90%
GPT-3.5	95.90%
GPT-4	**100%**

prevent the LLM's output from being disrupted, while still ensuring there are enough distractors available for selection. We ultimately set the candidate set size to 50.[14]

6.2 Evaluation of LLMs on English Vocabulary Test

We evaluate the LLM's performance on a vocabulary test because its ability to correctly answer these questions is fundamental to the distractor generation method. The LLM must be able to accurately distinguish between the correct answer and the distractors to ensure that the generated distractors are not mistakenly identified as valid answers, thus maintaining the requirement for a single correct answer in the question design. If the LLM incorrectly selects a distractor as the answer, it is highly likely that the distractor also fits the stem, thereby invalidating the generated vocabulary question. To assess this further, we conducted an experiment using the GSAT English test, where LLMs were presented with actual test questions and tasked with selecting the most appropriate answer from a set of well-designed plausible distractors. Table 6 shows the accuracy of various LLMs in answering English vocabulary questions. The results indicate that LLMs possess a strong capability in solving well-designed vocabulary problems. To sum up, we propose that the self-review mechanism may be effective, as the LLM can accurately select the correct answer without mistakenly choosing plausible distractors. To further validate this conclusion, additional investigation is required, which we will discuss in more detail in the following section.

6.3 Impact of Different Self-Review Methods

The main goal of a distractor validator is to ensure that the generated distractors do not serve as valid correct answers. Since LLMs possess the capability to solve a wide range of problems, there are numerous approaches to inference LLM to achieving this goal. To investigate the LLM's ability to distinguish between valid and invalid distractors using different queries, we designed an experiment. In this experiment, we examine the effectiveness of three different self-review methods with different prompts. We used `gpt-3.5-turbo-0125` as the distractor validator and utilize 193 GSAT English exam questions for evaluation. The core method we used to validate the appropriateness of a distractor involves presenting both the distractor and the question to the LLM, allowing it to evaluate the suitability using various assessment approaches. Ultimately, we selected the most effective evaluation method as the final self-review approach.

We separated the three golden distractors in each question into individual queries, resulting in a total of 579 queries for the LLM to evaluate the suitability of each distractor. By providing the golden distractor and target word, the distractor validator is expected to respond by recognizing the golden distractor as a valid option. The three prompts have the following query objectives:

Independent Suitability Judgment. In this prompt, we first instruct the LLM by informing it that it is an English teacher designing a vocabulary test for students. After providing the stem and target word, we ask the LLM whether the generated distractors are appropriate for this question. This design aims to test whether the LLM can independently judge if a distractor is suitable for the given question. The results indicate that the LLM confidently identified the golden distractor as a suitable option only 4 out of 579 times.

Semantic Consistency Check. In this version, we create two sentences by filling the incomplete stem with either the target word or the distractor. We first inform the LLM that the two provided sentences differ by only one word. Then, we ask LLM whether these two sentences convey the same meaning. This design aims to test whether the LLM can discern if the meaning changes when the target word is replaced with the distractor, thereby determining if the distractor alters the original meaning.

Since distractors typically convey a different meaning than the correct answer, especially when inserted into the stem, a valid distractor should result in a sentence with a different meaning compared to when the correct answer is used. The results show that in 397 out of 579 instances, the LLM identified that the golden distractor changed the meaning of the sentence. This approach proves to be more effective than the previous method, where the LLM was asked to directly assess the suitability of the distractors. This improvement may stem from the more explicit evaluation of meaning shifts, rather than relying solely on the model's inherent understanding of distractor quality. Additionally, the results suggest that the LLM lacks a clear internal understanding of the characteristics a distractor should possess, highlighting the need for more guided context-based evaluation.

Binary Choice Validation. According the results presented in Section 6.2, we found that LLMs excel at solving vocabulary test questions. Therefore, in this version, we formulate the task of evaluating distractor suitability into a binary choice. Specifically, the LLM is presented with both the correct answer and a distractor, and it is tasked with selecting the correct answer to fill in the stem, thereby determining the appropriateness of the distractor. Note that, different from standard single-choice vocabulary questions,

[14]Although ISSR may not be able to accommodate a large candidate set in a single run, one potential approach could involve using ISSR multiple times with a simple algorithm to iteratively select plausible distractors. This idea remains conceptual at this stage, and further exploration would be needed to assess its feasibility. Details of this potential algorithm are described in Appendix C.

we designed the prompt that allowed the LLM to consider both two options (i.e., the target word and the distractor) as possible answers to be filled in the stem, enabling it to recognize cases where either option could plausibly complete the question stem. If the LLM selected the distractor as a correct answer, this indicated that the distractor could potentially serve as an alternative valid answer when inserted into the stem. However, this is problematic because distractors are specifically designed to be incorrect, meaning they should never be considered valid answers when filling the stem. Hence, the intended outcome is that the model is able to consistently identify the target word as the correct answer from the two options and use it to fill in the stem. The results show that the LLM correctly selected the target word as the answer in 563 out of 579 instances. This indicates that the method is effective in filtering out unsuitable distractors, as it forces the model to make a clear distinction between the correct answer and the distractor. By placing the distractor in direct comparison with the target word, the model can better assess the appropriateness of each option, reducing the likelihood of selecting distractors that could be mistakenly viewed as fitting answers for the question, which would undermine the question's integrity. Among the 16 questions answered incorrectly, only two instances involved the LLM responding with "Both Are Good." This suggests that the LLM tends to select a single option as its output rather than indicating that both options are equally acceptable. Since the binary choice validation method performed the best, we adopted this approach as the self-review mechanism in ISSR. Detailed prompts used in this experiment can be found in Table 11.

6.4 Distractor Selection Evaluation

As discussed in Section 5.2, directly generating distractors using LLMs may not be the optimal solution due to their instability and limitations in generating a large number of distractors. However, it remains unclear whether LLMs possess the ability to effectively select appropriate distractors from a pool of distractor candidates. To this end, we examine the effectiveness of LLMs in selecting plausible distractors from a given candidate set. First, for each question stem and target word from the GSAT English vocabulary test, we generated 10 distractor candidates using the `BERT-base-uncased` model as the candidate generator. Next, we randomly replaced some of these candidates with the golden distractors from the actual exam, ensuring that each generated candidate set included appropriate distractors. We then provided the stem, target word, and candidate set to various LLMs, asking it to select the top 3 most suitable distractors. Table 7 shows the results selected by the LLMs, indicating that most LLMs were capable of selecting appropriate distractors from a well-curated candidate set. However, when distractors were not drawn from such specially designed candidate sets, the F-score drops significantly (presented in Table 2), indicating that the bottleneck lies in generating suitable candidate distractor sets. Therefore, improving the method of generating distractor candidate sets will be crucial to enhancing the overall system performance.

Table 7: LLM Distractor Selection Abilities

Model	Candidate Set Size	F1@3	NDCG@3
GPT-3.5	10	35.58%	58.60%
Llama3 8B	10	43.52%	71.15%
Llama3 70B	10	33.85%	56.65%

Table 8: Performance of ISSR Under Different Distractors Selection Count Per Round

Selection Size	F1@3	F1@10	NDCG@3	NDCG@10	NDCG@30
30	0.52%	1.19%	1.30%	3.20%	6.44%
10	0.69%	0.96%	1.55%	3.03%	7.49%
3	**1.55%**	**2.07%**	**3.57%**	**6.31%**	**9.82%**

6.5 Impact of Distractor Selection Count on LLM Performance

In our experimental results, we observed that LLMs possess the ability to select appropriate distractors from a candidate set. However, we also found that the number of distractors requested from the LLM during each selection round not only affects the quality of the chosen distractors but also impacts the overall validity of the results. In this subsection, we explore how varying the number of distractors selected per round influences the LLM's ability to effectively choose suitable distractors and maintain consistency in the selection process.

In the experimental setup, we designed a scenario where the LLM was tasked with selecting 30 distractors to be provided to teachers as suggestions for constructing exam questions. This setting aimed to simulate a realistic use case where a large number of candidate distractors are generated for teachers to review and choose from. Additionally, we varied the selection sizes to explore how different batch sizes affected the quality of the distractors.

Specifically, we experimented with selection sizes of 30, 10, and 3 distractors per round to analyze the LLM's performance in identifying suitable distractors across different conditions. As shown in Table 8, we experimented with different selection sizes, such as requesting 30 distractors in one round, 10 distractors in three rounds, and 3 distractors in ten rounds. The "selection size" indicates how many distractors the LLM was asked to generate in each round. From the results, we found that selecting 30 distractors at once does not yield the best results. In fact, smaller selection sizes, such as 3 or 10 distractors per round, lead to more effective outcomes, as reflected in higher F1 and NDCG scores. This demonstrates that requesting too many distractors in a single round can overwhelm the LLM and reduce the overall quality of its selections. Based on the experimental results, we determined that selecting three distractors at a time is the most effective strategy for the distractor selection process within the ISSR framework.

6.6 Human Evaluation of ISSR-Generated Questions

In our experiments, we demonstrated that ISSR outperforms other baseline models. However, it remains unclear whether the distractors generated by ISSR are genuinely plausible yet valid enough to effectively challenge examinees. To investigate this, we conducted an evaluation using a set of 30 questions from the GSAT exams. ISSR was used to generate distractors for each question, and 13 university students were invited to take the exam with these new distractors. All participants had sufficient English proficiency for the GSAT test and had shown strong performance in previous GSAT English assessments. The questions were selected based on their original pass rates. First, we sorted the questions by their pass rates and then divided them into three equal-sized groups, then we randomly select 10 questions from each group. The groups are organized as follows:

- **First Group:** This group includes the questions with the lowest pass rates, ranging from a minimum pass rate of 23% to a maximum of 54%. The standard deviation of pass rates within this group is 7.15.
- **Second Group:** The questions in this group have pass rates ranging from 54% to 65%, with a standard deviation of 3.27.
- **Third Group:** This group covers questions with the highest pass rates, from a minimum of 65% to a maximum of 87%. The standard deviation for this group is 5.30.

This approach allowed us to assess the performance of ISSR-generated distractors across questions of varying difficulty levels, providing insight into how effectively ISSR-generated distractors function in both easier and more challenging contexts.

Additionally, we asked the students to label any distractors they confidently identified as also be plausible as correct answers, aside from the most accurate option, which would render the question invalid. We also manually verified whether the generated distractors could indeed be correct answers for these labeled questions. According to our statistics, of the 90 generated distractor, a total of 8 were labeled as correct answers, potentially rendering the questions invalid. Among the distractors labeled as invalid, a student labeled the correct answer as an alternative plausible option, while selecting a distractor as the intended answer. This suggests that the distractors generated by ISSR support the question design, ensuring that only one option is the correct answer.

Table 9 shows the accuracy of the students. The "Challenging Questions" column denotes the number of questions that the students labeled as confusing due to the generated distractors. The results indicate that over half of the students achieved accuracy between 90 and 100. Within this group, 6 questions were identified as "plausible questions"—questions where the distractors could potentially mislead students. Given their strong background in English vocabulary, these students did not find most of the questions overly challenging, yet they still found 6 questions plausible. This suggests that ISSR is effective in generating distractors that are realistically capable of confusing students. Lastly, two examinees had accuracies between 60 and 70, yet did not annotate any distractors as confusing. This could be due to the difficulty they faced in recognizing vocabulary, which may have made it hard for them to discern the effectiveness of the distractors.

Table 9: Student Accuracy on ISSR Generated Distractors

Accuracy (%)	Students	Challenging Questions
90% ~ 100%	7	6
80% ~ 90%	3	4
70% ~ 80%	0	0
60% ~ 70%	2	0

Based on these observations, we conclude that ISSR-generated distractors are effective at subtly challenging students, particularly those with moderate to high proficiency.

7 CONCLUSION

Vocabulary acquisition is fundamental to mastering second languages, as a rich vocabulary enhances both comprehension and expression. Designing effective vocabulary tests is crucial in helping learners consolidate their understanding and identify gaps in their knowledge. In vocabulary tests, teachers often need to invest significant time and effort into generating suitable distractors—options that are misleading but cannot serve as correct answers. However, this process can be labor-intensive and prone to inconsistencies. Therefore, the ability to automatically generate plausible distractors is of great importance in reducing the workload for teachers and enhancing the quality of the tests. These automatically generated distractors help assess whether learners truly understand the meaning of the target words, rather than relying on mere guessing.

In this study, we present a framework named ISSR, aimed at assisting teachers in generating suitable distractors for vocabulary test design. To achieve this goal, we analyze vocabulary questions used in actual exams and examine the relationship between the target word and distractors. We also investigate the factors that make a distractor appealing to students and developed a set of predefined filtering rules to enhance distractor quality. Next, we explore the capabilities of LLMs in automatically generating distractors for vocabulary questions. Our findings indicate that LLMs perform better when selecting from a provided distractors candidates rather than generating distractors directly, due to the instability of LLMs and limitations in producing large numbers of distractors..

To tackle these issues, we propose a distractor selector, a module that leverages LLMs to select plausible distractors from the generated candidates. Finally, we introduced a distractor validator module with a self-review mechanism that leverages the LLM's ability to solve vocabulary questions. This mechanism filters out distractors that fit the stem but could lead to multiple valid answers, ultimately rendering the question invalid. Although we found that the proper use of LLMs can enhance the ability to automatically generate vocabulary question distractors, ISSR still has some drawbacks. Since ISSR leverages both the LLM and the BERT model simultaneously, it requires significantly more computing resources compared to similar work. Additionally, because the self-review mechanism involves converting questions into binary choices to individually verify the validity of candidate distractors, the distractor generation speed of ISSR is relatively slow. Addressing the limitations in computing resources and generation efficiency is left as our future work.

8 LIMITATIONS

The criteria derived from the GSAT dataset may not be universally applicable across all types of exams, as different standardized tests may vary significantly in structure and objectives. Our analysis and findings are specifically tailored to the characteristics of the GSAT, limiting their generalizability to other exam contexts. Additionally, the quality of the candidate set is highly dependent on the candidate generation mechanism. Since we employed existing methods, the effectiveness of later stages—distractor selection—is constrained by the initial quality of the candidate set, making this a bottleneck in the overall process. More advanced models or methods tailored to specific exam formats may be required to improve candidate generation.

Furthermore, while we have discussed the impact of leveraging in-context learning by providing demonstrations to LLMs, our exploration did not encompass all possible prompting strategies. There may be more effective prompting methods or alternative frameworks that could further enhance LLM performance in this task, and these have yet to be fully explored. This presents a significant opportunity for further optimization.

Future research could focus on incorporating additional generation constraints, refining prompt structures, and optimizing demonstration selection to guide LLMs in generating distractors that more closely resemble those created by educators. These steps may lead to further improvements in the quality and contextual relevance of the generated test items.

REFERENCES

[1] 1995.. *Collins Cobuild English dictionary.* HarperCollins.

[2] 2001. *Collins Cobuild English dictionary for advanced learners* (3rd ed. ed.). HarperCollins, Glasgow.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, null (mar 2003), 993–1022.

[4] Nathan Brake and Thomas Schaaf. 2024. Comparing Two Model Designs for Clinical Note Generation; Is an LLM a Useful Evaluator of Consistency?. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 352–363. https://doi.org/10.18653/v1/2024.findings-naacl.25

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[6] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[7] 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*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5835–5840. https://doi.org/10.18653/v1/2022.findings-emnlp.429

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[9] Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. *MIT Press google schola* 2 (1998), 678–686.

[10] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[11] John Brian Heaton. 1988. *Writing English language tests.* Longman.

[12] Shin'ichiro Ishikawa, Toshihiko Uemura, M Kaneda, Shinichi Shimizu, Naoki Sugimori, Yukio Tono, and M Murata. 2003. JACET8000: JACET list of 8000 basic words. *Tokyo: JACET* 3 (2003).

[13] Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints* (2024).

[14] Mehrdad Jalali, Yi Luo, Lachlan Caulfield, Eric Sauter, Alexei Nefedov, and Christof Wöll. 2024. Large language models in electronic laboratory notebooks: Transforming materials science research workflows. *Materials Today Communications* 40 (2024), 109801. https://doi.org/10.1016/j.mtcomm.2024.109801

[15] Shu Jiang and John Lee. 2017. Distractor Generation for Chinese Fill-in-the-blank Items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Joel Tetreault, Jill Burstein, Claudia Leacock, and Helen Yannakoudakis (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 143–148. https://doi.org/10.18653/v1/W17-5015

[16] Dongwoo Kim, Haixun Wang, and Alice Oh. 2013. Context-dependent conceptualization. In *Twenty-Third International Joint Conference on Artificial Intelligence.* Citeseer.

[17] 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*, Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 284–290. https://doi.org/10.18653/v1/W18-0533

[18] Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024. Let's Learn Step by Step: Enhancing In-Context Learning Ability with Curriculum Learning. *arXiv preprint arXiv:2402.10738* (2024).

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[21] 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*, Vol. 35. 4339–4347.

[22] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2655–2671. https://doi.org/10.18653/v1/2022.naacl-main.191

[23] Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic distractor generation for multiple-choice English vocabulary questions. *Research and practice in technology enhanced learning* 13, 1 (2018), 15.

[24] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Vol. 1. I–I. https://doi.org/10.1109/CVPR.2001.990517

[25] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 515–524. https://doi.org/10.1145/3077136.3080786

[26] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 968–979. https://doi.org/10.18653/v1/2023.emnlp-main.61

[27] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data.* 481–492.

[28] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale Cloze Test Dataset Created by Teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2344–2356. https://doi.org/10.18653/v1/D18-1257

[29] Carlos Alfredo Yorio. 1971. Some sources of reading problems for foreign-language learners 1. *Language learning* 21, 1 (1971), 107–115.

A PROMPTS USED IN THIS WORK

In this section, we present the prompts utilized in the ISSR framework and experiments. The prompt used for the distractor selector is detailed in Table 10. Table 11 presents the prompts employed for the self-review mechanism. Note that we have integrated "Binary Choice Validation" for determining distractor suitability within

Table 10: Prompt used in Distractor Selector

Input: **Original Sentence** Posters of the local rock band were displayed in store windows to promote the sale of their _____ tickets. **Target Word** concert **Candidate Pool** "sports", "proper", "regular", "personal", "clothes", "favorite", "traffic", "traditional", "valuable", "available", "travel", "necessary", "fashionable", "record", "official", "final", "usual", "clothing", "educational", "fashion", "journey" pick three distractors from **Candidate Pool** for stem given in Original Sentence, response each distractors per line, and starts with enumerate number.
Output: 1. journey 2. traffic 3. record

Table 11: Prompt used in Self Review

[Binary Choice Validation for Distractor Suitability] **Input:** Imagine you are a high school student that studying english, and you are answering question given below: The following is a vocabulary test that requires selecting one answer from given options to fill in the blank. Please select the option that fit the context best from below, response with the correct option directly, if you think both options are suitable for the context, response with "BOTH ARE GOOD". Question: The newcomer speaks with a strong Irish _____; he must be from Ireland. options: identity accent
[Independent Suitability Judgment for Distractor Validation] **Input:**Imagine you are a english teacher that designing a vocabulary test to a second language learner, and you came up with a distrctor candidate "identity". Qustion: The newcomer speaks with a strong Irish _____; he must be from Ireland. Correct answer: accent Distractor candidate: identity The criteria for question creation are as follows: 1. The length difference between the answer and the distractor should not exceed 2 characters. 2. The answer and the distractor should share the same part of speech. 3. The difficulty levels between the answer and distractor should be closely matched Do you think whether word "identity" is a good distractor or not? Response with Yes or No only.
[Semantic Consistency Check for Distractor Validation] You will now see two sentences with only one word difference between them: Sentence 1: The newcomer speaks with a strong Irish identity; he must be from Ireland. Sentence 2: The newcomer speaks with a strong Irish accent; he must be from Ireland. Do these two sentences have the same meaning? Please respond with 'Yes' or 'No' only

our ISSR framework, as this method effectively filters out invalid distractors with the highest accuracy.

B EXAMPLES OF DIRECTLY GENERATE DISTRACTORS WITH LLMS

As discussed in Section 5.2, directly inferring from `gpt-3.5-turbo-0125` may result in repetitive content when generating a large number of distractors. However, it remains unclear whether this issue arises in other LLMs. In this section, we examine the following two issues: (1) The tendency of `gpt-3.5-turbo-0125` to generate repetitive content: We investigate patterns and underlying reasons for similar or repetitive distractors when using `gpt-3.5-turbo-0125` to generate multiple-choice questions. (2) Whether other LLMs also exhibit a similar tendency to generate repetitive distractors.

Tendency of GPT-3.5-turbo to Generate Repetitive Distractors We explore two intuitive methods for generating 30 distractors using LLMs: (1) Instructing `GPT-3.5-turbo-0125` to generate all 30 distractors in a single round, and (2) Requesting `GPT-3.5-turbo-1106` to generate 3 distractors per round, with explicit instructions to avoid repeating any previously generated distractors across rounds.

Table 12 and 13 present the results of the first and second methods, respectively. The issue with the first method is not only that `gpt-3.5-turbo-0125` generates repetitive content but also that it begins producing phrases rather than single vocabulary items, which are inappropriate for actual exams. This issue did not occur with the second method.

As shown in Table 12, `gpt-3.5-turbo-1106` initially generates distractors successfully and avoids repeating restricted distractors. However, after several rounds, `gpt-3.5-turbo-1106`'s output becomes unstable, and it starts to repeat distractors, even though they were restricted in the prompt.

We also found that in both methods, `gpt-3.5-turbo-0125` attempts to generate terms that fit the context, which may compromise the validity of the vocabulary questions. While this issue might be mitigated by using different prompts or leveraging in-context learning, it remains uncertain whether the LLM can consistently avoid generating invalid distractors. Thus, we conclude that directly generating distractors via `gpt-3.5-turbo-0125` is ineffective due to the challenges elaborated above.

Exploring Whether Other LLMs Exhibit a Tendency to Generate Repetitive Distractors We further examined the performance of Llama3 8B and Llama3 70B in generating a large number of distractors, employing the same method used with `GPT-3.5-turbo`. Table 14 presents an example in which Llama3 70B successfully generates a large number of distractors in a single round, while Table 15 shows the results of generating a smaller number of distractors per round, repeated until the desired number of distractors is

Table 12: Example of Directly Inference GPT-3.5 to Generate Large Quantity of Distractors with Small Generate Count Per Round

---

Input: **Original Sentence** → `[Stem]`
The newcomer speaks with a strong Irish _____; he must be from Ireland.
Target Word → `[Target Word]`
accent
generate 3 distractors for stem given in Original Sentence, restrain your output in following format given below (for example, your generated 3 distractors are: apple, banana, orange, ...)
1. apple
2. banana
3. orange → `[Request]`

Output:1. brogue 2. lilt 3. twang

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, twang, lilt

`[Same Request]`
Output:1. dialect 2. intonation 3. inflection

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
dialect, lilt, brogue, inflection, intonation, twang

`[Same Request]`
Output:1. lilt 2. twang 3. inflection

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
dialect, lilt, brogue, intonation, inflection, twang

`[Same Request]`
Output:1. tone 2. speech 3. drawl

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
dialect, speech, tone, lilt, brogue, drawl, inflection, intonation, twang

`[Same Request]`
Output:1. lilt 2. brogue 3. twang

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
dialect, speech, tone, lilt, brogue, drawl, inflection, intonation, twang

`[Same Request]`
Output:1. lilt 2. brogue 3. twang

---

achieved. These examples demonstrate that Llama3 70B effectively performs the task, producing plausible distractors that meet the requirements of the exam.

On the other hand, Llama3 8B struggles in this task. As shown in Table **??**, attempting to generate a large number of distractors

Table 13: Example of Directly Inference GPT-3.5 to Generate Large Quantity of Distractors in Single Round

---

Input: **Original Sentence**
In the cross-lake swimming race, a boat will be _____ in case of an emergency.
Target Word
standingby
Words to avoid picking
brogue, twang, lilt
generate 30 distractors for stem given in Original Sentence, restrain your output in following format given below (for example, your first three generated distractors are: apple, banana, orange)
1. apple
2. banana
3. orange

Output:1. ready
2. available
3. prepared
4. on standby
5. waiting
6. on call
7. on alert
8. on hand
9. at the ready
10. on standby
11. on standby
12. on standby
13. on standby
14. on standby
15. on standby
16. on standby
17. on standby
18. on standby
19. on standby
20. on standby
21. on standby
22. on standby
23. on standby
24. on standby
25. on standby
26. on standby
27. on standby
28. on standby
29. on standby
30. on standby

---

in a single round often results in repeating output. Similarly, Table 17 shows that even when generating a small number of distractors per round, Llama3 8B encounters similar challenges to `gpt-3.5-turbo-1106` in producing consistent, contextually relevant distractors.

Although some models, like Llama3 70B, demonstrate reliable performance in generating multiple distractors with both method, the results suggest the performance is strongly influenced by the model's size. To enhance the stability and quality of distractor generation across different LLMs, we propose an alternative approach called the distractor selector in ISSR: instead of directly generating distractors, the model could select plausible distractors from

Table 14: Example of Directly Inference LLAMA3.1-70B to Generate Large Quantity of Distractors in Single Round

Input: **Original Sentence**
In the cross-lake swimming race, a boat will be _____ in case of an emergency.
Target Word
standingby
Words to avoid picking
brogue, twang, lilt
generate 30 distractors for stem given in Original Sentence, restrain your output in following format given below (for example, your first three generated distractors are: apple, banana, orange)
1. apple
2. banana
3. orange
Output: Here are 30 distractors for the target word "accent" in the given sentence: 1. dialect
2. brogue
3. tone
4. inflection
5. pitch
6. timbre
7. voice
8. pronunciation
9. lilt
10. idioms
11. slang
12. drawl
13. twang
14. cadence
15. rhythm
16. emphasis
17. mannerism
18. speech
19. articulation
20. enunciation
21. modulation
22. expression
23. phraseology
24. parlance
25. vocabulary
26. elocution
27. diction
28. syntax
29. utterance
30. intonation

Table 15: Example of Directly Inference LLAMA3.1-70B to Generate Large Quantity of Distractors with Small Generate Count Per Round

Input: **Original Sentence** ⟶ `[Stem]`
The newcomer speaks with a strong Irish _____; he must be from Ireland.
Target Word ⟶ `[Target Word]`
accent
generate 3 distractors for stem given in Original Sentence, restrain your output in following format given below (for example, your generated 3 distractors are: apple, banana, orange, ...)
1. apple
2. banana
3. orange ⟶ `[Request]`
Output:1. brogue 2. dialect 3.tone

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, dialect, tone

`[Same Request]`
Output:1. inflection 2. lilt 3. twang

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, dialect, tone, inflection, lilt, twang

`[Same Request]`
Output:1. pronunciation 2. cadence 3. drawl

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, dialect, tone, inflection, lilt, twang, pronunciation, cadence, drawl

`[Same Request]`
Output:1. timbre 2. pitch 3. intonation

a predefined candidate set, potentially improving consistency and accuracy.

C ALGORITHM FOR SELECT DISTRACTORS FROM NUMEROUS CANDIDATES

Although ISSR is capable of selecting appropriate distractors from a sufficient set of candidates, there may be circumstances where an excessively large number of distractor candidates are provided. In this appendix, we discuss a potential solution to address ISSR's limitations in handling large candidate sets through a simple algorithm inspired by median of medians. The algorithm works by partitioning the large distractor set into K groups, where $K = \lceil \frac{|C|}{30} \rceil$ and $|C|$ represents the total number of distractor candidates. After partitioning, distractor selector in ISSR is applied to each group individually, selecting the top N distractors from each group, where $N = \lfloor \frac{30}{K} \rfloor$. Once the top N distractors from each group are selected, they are merged to form a new candidate set. Distractor selector in ISSR is then invoked again on this merged set to finalize the selection of distractors. This algorithm allows us to maintain the performance of ISSR even when dealing with a large number of distractor candidates, while minimizing the loss of potentially good candidates.

Table 17: Example of Directly Inference LLAMA3.1-8B to Generate Large Quantity of Distractors with Small Generate Count Per Round

Input: **Original Sentence** → `[Stem]`
The newcomer speaks with a strong Irish _____; he must be from Ireland.
Target Word → `[Target Word]`
accent
generate 3 distractors for stem given in Original Sentence, restrain your output in following format given below (for example, your generated 3 distractors are: apple, banana, orange, ...)
1. apple
2. banana
3. orange → `[Request]`
Output:1. brogue 2. twang 3. dialect

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, dialect, twang `[Same Request]`
Output:1. intonation 2. cadence 3. modulation

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, dialect, twang, intonation, cadence, modulation

`[Same Request]`
Output:1. brogue 2. dialect 3. cadence

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, intonation, dialect, twang, cadence, modulation

`[Same Request]`
Output:1. pronunciation 2. cadence 3. drawl

Input: `[Same Stem]`

`[Same Target Word]`
Words to avoid picking
brogue, intonation, dialect, twang, cadence, modulation, pronunciation, drawl

`[Same Request]`
Output:1. drawl 2. brogue 3. cadence