Iterative Selection with Self-Review for Vocabulary Test Distractor Generation

Yu-Cheng Liu liu2022113.cs11@nycu.edu.tw
An-Zi Yen azyen@nycu.edu.tw
Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Editors: Hung-yi Lee and Tongliang Liu

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 frequently produce distractors that risk invalidating the question by introducing multiple correct options. In this study, we focus on English vocabulary questions. We analyze how teachers design test items to gain insights into distractor selection strategies. Additionally, we identify 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 makes use of an LLM-based self-review mechanism to ensure that the distractors remain valid while offering diverse options. ISSR aims to assist educators by providing an LLM-based tool that allows them to efficiently design pedagogically sound distractors through natural language instructions. 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

Expanding one's vocabulary is a key factor in enhancing overall language comprehension (Yorio, 1971). Given the importance of vocabulary acquisition, effective assessment methods are essential. A common approach is gap-filling multiple-choice questions, which test both word meaning and contextual usage. For example, in the sentence "Posters of the local rock band were displayed in store windows to promote the sale of their ___ tickets," the missing word "concert" must be chosen from four options. Such vocabulary items are a staple of standardized English exams in Asia. These exams include the General Scholastic Ability Test (GSAT),[1] Jitsuyo Eigo Gino Kentei (EIKEN),[2] the Advanced Subjects Test (AST),[3] and the Test of English Proficiency developed by Seoul National University

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

(TEPS).[4] These indicate the critical importance of well-designed vocabulary questions in assessing language proficiency.

Effective distractors force examinees to discern subtle meaning differences rather than guess, leading to a more accurate assessment of language proficiency. However, manual design of distractors is labor-intensive. In this work, we seek to explore effective approaches for the automatic design of English vocabulary test items that accurately assess learners' vocabulary comprehension. The automatic generation of distractors based on a given stem and target word has gained research attention, providing teachers with selection options. Susanti et al. (2018) extract candidates from lexical databases and word lists, measuring semantic relatedness, but are limited by fixed lists, reducing diversity. Liang et al. (2018) train a model on manual features to generate exam-like distractors, while Chiang et al. (2022) combine neural networks and predefined rules for diverse, effective distractors. Although prior methods (Liang et al., 2018; Chiang et al., 2022) show promise, they require extensive training data, often infeasible for specialized vocabulary tests, highlighting the need for approaches independent of large-scale data or predefined dictionaries.

Recent LLMs have demonstrated powerful semantic understanding and language generation abilities. Given these advantages, perhaps LLMs could be leveraged to address the aforementioned challenges. However, it remains unclear whether LLMs can effectively perform the task of distractor generation. This lead to the first research question (**RQ1**): **Can LLMs be directly utilized to generate distractors?**

Although LLMs excel in language generation, generating effective distractors requires precise control over their similarity to the correct answer. Distractors must be misleading yet distinct, avoiding excessive alignment with the stem to prevent multiple correct answers. Additionally, they must be contextually appropriate and free from logical inconsistencies.

Another critical factor is the difficulty level of distractors, which must balance being misleading without being overly simple or difficult. This balance is essential for effectively distinguishing learners of different proficiency levels. Hence, we raise the second research question (**RQ2**): **How can we ensure that the generated distractors remain valid and do not introduce ambiguity or multiple correct answers?**

To address the research question, this paper examines English vocabulary test items from an Asian university entrance exam, using the GSAT as a case study. The GSAT was selected for its detailed records, including option selection rates across students of different proficiency levels. Its test items cover a broad vocabulary range and incorporate distractors of varying difficulty, providing a comprehensive framework for evaluating vocabulary mastery. By analyzing vocabulary range, distractor correctness, and common student errors, we assess the effectiveness and challenges of existing test items. We propose the Iterative Selection with Self-Review (ISSR) framework, which consists of three modules: a candidate generator, a distractor selector, and a distractor validator. ISSR employs a pretrained language model (PLM) to generate contextually relevant distractors and integrates an LLM-based self-review mechanism to ensure that no distractor is mistakenly a valid answer. This approach reduces manual effort while enhancing the diversity and accuracy of test items, providing a more effective assessment of students' vocabulary comprehension. Additionally, ISSR offers intuitive natural language control, allowing teachers to customize test content

4. https://en.teps.or.kr/about_teps.html

without technical expertise. Through natural language instructions (prompts), educators can specify requirements such as distractor difficulty or contextual appropriateness, and the system automatically generates suitable options. ISSR's independence from fine-tuning data enhances efficiency and usability, making it a scalable and flexible tool for diverse assessment needs. 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 the generation of contextually appropriate distractors that are valid and avoid ambiguity.
- We propose ISSR, leveraging a PLM and LLM-based self-review to enhance distractor quality, reduce manual effort, and improve test adaptability.[5]
- Experimental results show that ISSR-generated distractors align with exam requirements, making it a practical tool for teacher collaboration. The self-review mechanism further ensures reliable distractor selection.

2. Related Work

Recent studies on automatic distractor generation often adopt a two-step approach: (1) generating distractor candidates and (2) ranking them to select the most plausible ones.

2.1. Distractor Candidate Generation

This step involves generating a broad set of distractor candidates to be further filtered in the next step. Susanti et al. (2018) propose a method to obtain and rank distractors from lexical databases and dictionaries. Their approach generates distractors using the target word, stem, and correct answer, ensuring semantic similarity while maintaining distinction. Candidates are retrieved from the text passage based on part of speech, tense, and sibling words from WordNet (Fellbaum, 1998) and JACET8000 (Ishikawa et al., 2003). Observing that human-created distractors share a similar difficulty level with the correct answer, they select words accordingly, expanding to Merriam-Webster if candidates are insufficient.

Previous studies mainly relied on lexical databases and dictionaries for distractor selection. Chiang et al. (2022) explore using pretrained language models (PLMs) for distractor generation, focusing on the CLOTH dataset (Xie et al., 2018), which comprises teacher-created cloze-type questions assessing vocabulary and grammar. They fine-tune PLMs to generate plausible distractors for the question stem rather than correct answers, leveraging the models' ability to predict masked words or complete sentences.

2.2. Distractor Candidate Scoring

In this step, candidates are ranked based on their suitability as effective distractors to select the most competitive options. Susanti et al. (2018) emphasize that effective distractors should be semantically similar to the target word yet distinct from the correct answer, using a ranking formula based on semantic similarity and collocation with GloVe word vectors (Pennington et al., 2014) and NLTK. Ren and Zhu (2021) advance the process by transforming candidates into 33-dimensional feature vectors to train ranking models, including AdaBoost (Freund and Schapire, 1997), LambdaMART (Burges, 2010), and other

5. https://github.com/NYCU-NLP-Lab/ISSR

rankers, enabling sophisticated selection strategies. Liang et al. (2018) propose NN-based and feature-based models for selecting plausible distractors, employing classifiers like logistic regression and random forests (Breiman, 2001) alongside adversarial training frameworks (IR-GAN) (Wang et al., 2017) and cascaded learning (Viola and Jones, 2001) for efficient filtering and ranking. Chiang et al. (2022) introduce a scoring mechanism that combines PLM confidence scores with semantic similarity from contextual and word embeddings, assigning weights to scoring methods to identify the top k distractors. In summary, our work builds on prior methods for distractor generation and scoring but diverges by incorporating LLMs to address issues such as dictionary reliance, large training data needs, and filtering challenges. By integrating LLMs, we seek to enhance flexibility and precision while retaining the strengths of earlier approaches.

3. Data Analysis

We selected questions from the GSAT English exam as our dataset, with past exam questions available on the College Entrance Examination Center (CEEC) website.[6] We gathered exams from 2004 to 2016, comprising a total of 195 questions. To develop a method for generating English vocabulary questions aligned with teacher-designed tests and to study effective distractor traits, we analyze question characteristics.

3.1. Analysis of CEEC wordlist

The CEEC has published a Senior High School English Wordlist,[7] detailing the vocabulary that Taiwanese high school students should understand before taking the GSAT and AST English tests. To analyze vocabulary difficulty, we reference the CEEC Senior High School English Wordlist, which categorizes words into six levels based on frequency in the Cobuild English Dictionary,[8] with adjustments for derivatives. To examine the difficulty distribution of vocabulary in the GSAT English exam, we analyzed the difficulty levels of target words based on the CEEC word list. Figure 1 shows that most words fall into levels 3 and 4, with approximately 60 and 80 words, respectively, while fewer words appear at levels 1, 2, 5, and 6. This suggests the exam primarily assesses moderately difficult vocabulary.

3.2. Relation Between Distractors and Answer

We examined GSAT distractor difficulty to evaluate their effectiveness in assessing students' vocabulary skills. Figure 2 shows the difference between the distractors and the target word. The x-axis represents the difficulty levels of the target words, and the y-axis represents the number of words at each difficulty level. In our analysis of the GSAT English exam, we found that most distractors differ from the target word by 0 to 2 difficulty levels, with only a few exceeding 3 or 4 levels. Most distractors are at most one level higher than the correct answer, likely due to word frequency grouping. To align with real exam patterns, our framework integrates this difficulty difference when generating distractors.

6. https://www.ceec.edu.tw/xmfile?xsmsid=0J052424829869345634
7. https://www.ceec.edu.tw/SourceUse/ce37/ce37.htm
8. https://www.collinsdictionary.com/dictionary/english

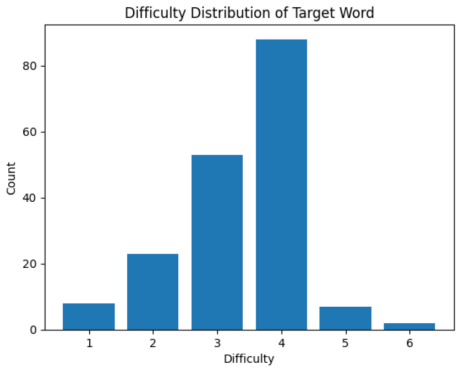

Figure 1: Difficulty distribution of target word

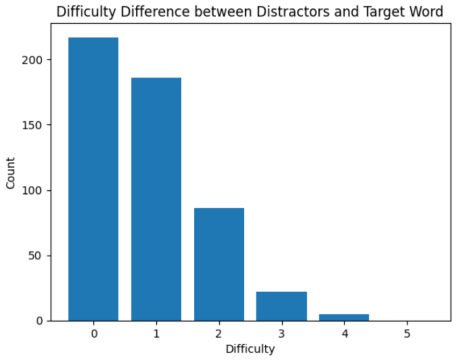

Figure 2: Difficulty difference between distractors and answers

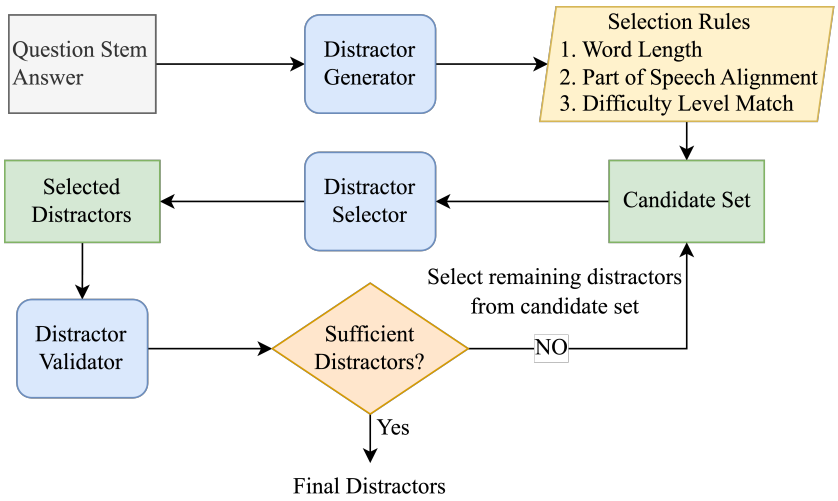

Figure 3: ISSR framework

4. Iterative Selection with Self-Review

We propose the ISSR framework to assist teachers in designing English vocabulary tests by generating distractors. As shown in Figure 3, it consists of a candidate generator, a distractor selector, and a distractor validator.

4.1. Candidate Generator

Given a question stem s and target word w, this module generates candidate distractors for selection. Note that direct distractor generation using LLMs led to instability, repetition, and inconsistent difficulty levels. Detailed discussion is in Section 6.1. To balance efficiency and specificity, we employ the BERT-based model CDGP-CSG $\mathcal{G}_{\text{CDGP}}$, fine-tuned on the CLOTH dataset Chiang et al. (2022). The initial distractors are obtained as $D = \mathcal{G}_{\text{CDGP}}(s, w; \theta)$, where θ denotes the fine-tuned parameters. Prior analysis in Sec-

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 1: Prompt used in distractor selector

tion 3 shows GSAT distractors typically differ by at most one difficulty level from the correct answer, a pattern our filtering enforces. To ensure distractor quality, we further apply the following criteria Heaton (1988) to construct the candidate set $C = \{c_1, \ldots, c_n\}$: (1) The length difference between the answer and distractor should not exceed two characters. (2) The answer and distractor must share the same part of speech. (3) The difficulty levels of the answer and distractor should be closely matched.

Distractors are reinserted into the stem for part-of-speech verification (NLTK), and difficulty levels are checked via the CEEC wordlist. This hybrid approach combines PLM flexibility with controlled filtering, maintaining coherence in iterative selection and ensuring GSAT exam alignment.

4.2. Distractor Selector

We utilize in-context learning to automatically select appropriate distractors based on instructions without requiring additional training. An LLM-based distractor selector is incorporated into ISSR to select the top n distractors from the candidate set. The distractor selector prompt is presented in Table 1. Specifically, we employ an LLM-based distractor selector $\mathcal{M}_{\text{sel}}$ to pick the top n distractors from C, i.e., $\{d_1, \ldots, d_n\} = \mathcal{M}_{\text{sel}}(S_{sel}, C)$, where each $d_i \in C$ and S_{sel} is the selector prompt. Next, according to the selection rules, a post-processing filter f removes any d_i that violates English vocabulary exam constraints, yielding the selected distractor set $D^{'} = f(\{d_1, \ldots, d_m\})$.

4.3. Distractor Validator

Sometimes, the candidate distractors generated in the previous steps could also serve as plausible answers, potentially resulting in multiple correct answers in the question. As the test design requires only one correct answer among the four options, further filtering ensures that distractors cannot function as valid correct answers. Thus, we propose a self-review mechanism to assess the validity of each distractor. New questions are formulated by pairing each distractor with the correct answer as a binary choice, allowing the LLM to evaluate each distractor's suitability. If the LLM selects the distractor as the correct answer, we take this to indicate that the distractor is unsuitable, as it yields a question with multiple valid answers. If the number of qualified distractors is insufficient, all distractors used in this

[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 distractor 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

Table 2: Prompt used in self review

round are removed from the candidate set, and the selector is rerun to obtain additional distractors.

Concretely, for every $d \in D^{'}$ we form a binary-choice question and use an LLM $\mathcal{M}_{\text{rev}}$ to predict $y = \mathcal{M}_{\text{rev}}(w, d_i)$. If $y = d_i$, we regard d_i as invalid since the model treats it as a correct answer, leaving multiple valid options; otherwise, d_i remains valid. The prompt template for $\mathcal{M}_{\text{rev}}$ presents in Table 2, and alternative self-review variants are discussed in Section 6.2. After filtering, let V denote the set of valid distractors. If $|V| < k$, we remove both valid and invalid elements from C and re-invoke the distractor selector $\mathcal{M}_{\text{sel}}$ to obtain additional candidates until $|V| \geq k$.

To sum up, our approach follows an Iterative Selection process rather than direct distractor generation using LLMs. First, the Candidate Generator produces potential distractors and applies filtering rules. Then, the Distractor Selector picks the most appropriate ones from the candidate pool, reducing instability and difficulty mismatches. Finally, the Distractor Validator ensures that distractors do not serve as unintended correct answers through a self-review mechanism. If a distractor is deemed unsuitable, the system reselects from the candidate pool and re-evaluates until a valid set is obtained, ensuring alignment with exam requirements.

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%
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%

Table 3: Distractor generation performance

5. Experiments

5.1. Experimental setup

We utilized English vocabulary questions from the GSAT spanning from 2006 to 2018. The dataset consisted of 195 questions. We applied 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, and the remaining 193 questions served as the test set. In the zero-shot prompting setting, the same 193 questions were directly used as the test set without prior demonstrations. To ensure the LLM returned the desired distractors, we structured prompts with explicit formatting instructions. The temperature for all LLMs was set to 0.7. We compared the ISSR framework against the following baselines: CDGP and GPT-3.5 (Ouyang et al., 2022). The GPT-3.5 model was `gpt-3.5-turbo-0125`.[9]

5.2. Experimental Results

Table 3 compares the results of ISSR to that of 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 Normalized Discounted Cumulative Gain (NDCG) were adopted as the evaluation metrics. NDCG@k evaluates ranking quality by giving higher importance to distractors that closely match teacher-designed ones at higher positions in the list. It compares the generated ranking with an ideal ranking, where all teacher-designed distractors are ranked at the top. We assign equal importance to all teacher-designed distractors in our evaluation.

As shown in Table 3, ISSR not only surpasses the direct use of GPT-3.5 for distractor generation, but also outperforms CDGP. This improvement stems from the ability of LLMs to select more contextually appropriate distractors rather than merely generating words that fit the sentence structure. This finding highlights the importance of a well-curated candidate set in generating high-quality distractors. Moreover, ISSR proves more effective in producing distractors that closely match those in actual exams, providing educators with a broader selection of viable options. Furthermore, Table 3 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.

Further discussion and examples comparing ISSR to direct generation methods can be found in Section 6.1. We found that when GPT-3.5 is tasked with generating a large set of distractors in a single round, it often produces repetitive content. Thus, we experimented

9. While our experiments do not include newer models due to the time required for evaluation and analysis, our framework remains compatible with different LLMs.

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%**

Table 4: Performance using different candidate generators

with generating a smaller number of distractors over multiple rounds, which partially reduced repetition.

Note that the performance of all methods is lower because the models select distractors from the output of the distractor generator, which may not include suitable distractors. However, without filtering candidates through the distractor generator, the distractor selector would need to choose from thousands of candidates.

Performance of different candidate generators. We tested the impact when selecting different sources as the candidate generator on ISSR. The comparison involves three settings: (1) no candidate generator, denoted as None, where distractors are directly generated using GPT-3.5 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. 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 was applied uniformly across all candidate generators.

Table 4 presents the results when using different models as the candidate generators. The results show that CDGP-CSG outperforms BERT-base. The standard BERT-based candidate generator performs weaker, as BERT's pretraining on masked token and next sentence prediction leads it to generate contextually fitting but suboptimal distractors. In contrast, the CDGP-CSG model, fine-tuned on the CLOTH dataset, learns to generate plausible distractors based on question stems rather than just fitting the context, despite differences between CLOTH and GSAT.

Results Using Different LLMs for Distractor Selection. The comparison involved three different models with varying parameter sizes: (1) GPT-3.5, (2) `Llama3-8B` (Dubey et al., 2024), and (3) `Llama3-70B`. For each model, both zero-shot and few-shot settings were evaluated; the CDGP-CSG model was used as the candidate generator.

Table 5 shows the results using different LLMs to select distractors: GPT-3.5 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. Additionally, the zero-shot slightly outperforms few-shot, suggesting that the LLM is capable of effectively selecting distractors without requiring prior demonstrations. Based on these findings, we adopt GPT-3.5 with zero-shot prompting in ISSR.

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%

Table 5: Results using different LLMs for distractor selection

6. Analysis and Discussion

6.1. Direct Generation of Distractors Using LLMs

We conduct an exploratory evaluation of LLMs' ability to generate multiple-choice distractors, focusing on qualitative patterns observed in a limited set of test cases. In this section, we use one test question to demonstrate the challenges of directly generating distractors with LLMs. The actual test question designed by teachers is as follows:

> In the cross-lake swimming race, a boat will be _____ in case of an emergency.
> (A) standing by (B) turning on (C) getting on (D) running down

As shown in Table 6, the results indicate that `GPT-3.5-turbo-0125` frequently produces repetitive outputs. For example, when given the sentence "In the cross-lake swimming race, a boat will be ___ in case of an emergency." with the target word "standing by," the model generated "on standby" more than 15 times in a single round. In Table 7, Llama3-70B demonstrated improved diversity but struggled with semantic distinction and contextual fit. Some distractors, like "waiting," "ready," and "prepared," were too close in meaning to "stand by." If distractors are too semantically similar to the correct answer, they may introduce ambiguity, potentially making the question invalid by allowing multiple plausible answers. This risks violating the single-correct-answer principle in multiple-choice tests and affecting the validity of the assessment. While others, such as "sailing" and "rowing," were contextually inappropriate. This highlights the challenge of generating distractors that are both distinct and relevant. Table 8 shows that Llama3-8B struggled to maintain contextual relevance, producing distractors such as "shocked" and "dazed," which are semantically unrelated to the target word "standing by." When distractors are too unrelated to the correct answer, they fail to serve their intended function of creating plausible but incorrect alternatives. The question becomes easier than intended, diminishing its ability to accurately assess whether students truly understand the target concept. These findings suggest that while larger models generally enhance output diversity, no model guarantees consistently appropriate distractors. To mitigate these issues, we propose a distractor selection approach from predefined candidate sets to enhance stability and accuracy.

6.2. Impact of Different Self-Review Methods

We evaluated three self-review methods for LLM-based distractor validation using GPT-3.5 and 193 GSAT English exam questions, yielding 579 evaluation queries. The goal was to determine the most effective approach for identifying distractors that are not valid answers.

Input: **Original Sentence**		
In the cross-lake swimming race, a boat will be _____ in case of an emergency.		
Target Word		
standing by		
Please 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	11. on standby	21. on standby
2. available	12. on standby	22. on standby
3. prepared	13. on standby	23. on standby
4. on standby	14. on standby	24. on standby
5. waiting	15. on standby	25. on standby
6. on call	16. on standby	26. on standby
7. on alert	17. on standby	27. on standby
8. on hand	18. on standby	28. on standby
9. at the ready	19. on standby	29. on standby
10. on standby	20. on standby	30. on standby

Table 6: Directly using GPT-3.5 to generate 30 distractors in a single round

Same prompt as Table 6		
Output: Here are 30 distractors for the target word "standing by" in the given sentence:		
1. waiting	11. hovering	21. patrolling
2. floating	12. lingering	22. escorting
3. nearby	13. loitering	23. accompanying
4. available	14. idling	24. following
5. ready	15. drifting	25. tracking
6. prepared	16. cruising	26. monitoring
7. anchored	17. sailing	27. observing
8. moored	18. rowing	28. watching
9. stationed	19. paddling	29. guarding
10. parked	20. surfing	30. fishing

Table 7: Directly using LLAMA3-70B to generate 30 distractors in a single round

(1) Independent Suitability Judgment: The LLM was prompted as an English teacher to assess whether distractors were appropriate. It correctly identified the golden distractor as suitable in only 0.69% (4/579) of cases, indicating a lack of intrinsic understanding of distractor quality.

(2) Semantic Consistency Check: This approach tested whether replacing the target word with the distractor altered the sentence's meaning. The LLM was shown two sentences—one with the correct answer and one with the distractor—and asked whether they conveyed the same meaning. Since valid distractors should change the meaning, a correct response meant recognizing this difference. The LLM succeeded in 68.57% (397/579) of cases, performing better than direct suitability judgment. However, results suggest the model lacks a strong internal understanding of what makes a good distractor, requiring more structured evaluation.

(3) Binary Choice Validation: this method framed distractor validation as a binary-choice task. The LLM was given both the correct answer and a distractor and asked to select the most suitable option to complete the sentence. If it selected the distractor, it implied the

Same prompt as Table 6		
Output: To generate distractors for the given sentence, we'll consider words that fit grammatically and contextually. The target word is "standing by," which means ready and prepared to act in case of an emergency.		
1. anchored	11. grounded	21. keeled
2. drifting	12. marooned	22. floundering
3. floating	13. castaway	23. foundering
4. moored	14. wrecked	24. listing
5. docked	15. sunken	25. tilting
6. beached	16. listing	26. aghast
7. stranded	17. tilting	27. shocked
8. afloat	18. keeling	28. stunned
9. buoyed	19. capsized	29. numbed
10. submerged	20. overturned	30. dazed

Table 8: Directly using LLAMA3-8B to generate 30 distractors in a single round

distractor could be a valid answer, which is undesirable. The LLM correctly chose the target word in 97.24% (563/579) of cases, making this the most effective approach. Notably, in the 16 incorrect cases, only two responses indicated both options were acceptable, suggesting a strong preference for a single correct choice.

Given its superior accuracy, binary choice validation was adopted as the self-review mechanism in ISSR.

6.3. Impact of Candidate Set Size

Perhaps it is natural to think that providing more distractor candidates to the distractor selector would 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 can 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 that do not appear in the distractor candidate set also increases.

To more precisely analyze the impact of the candidate set size on the LLM's ability to effectively select distractors, we conducted an experiment to determine the optimal candidate set size. We utilized ISSR and varied the number of candidate distractors provided to the distractor selector. Specifically, we extracted the stem and target word from the original teacher-designed test questions, and constructed inputs using candidate sets of varying sizes generated by the distractor generator. We then evaluated whether the distractor selector could accurately identify distractors from each candidate set.

Table 9 shows a negative correlation between candidate set size and the LLM's ability to select teacher-designed distractors. The LLM was GPT-3.5. The success rate represents the percentage of cases in which the LLM correctly selects all three intended distractors from the candidate pool. When the candidate set is small, the LLM reliably identifies the intended distractors, but as the set size increases, selection accuracy declines. This suggests that a larger pool may introduce ambiguity, causing the model to infer unintended relationships among candidates rather than strictly following the selection criteria. To balance accuracy and diversity, we set the candidate size to 50.

Candidate Size	Success Rate
300	90.67%
200	92.75%
100	97.93%
50	**98.79%**

Table 9: Relationship between candidate size and successful distractor selection rate

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

Table 10: Student accuracy on ISSR generated distractors

7. Human Evaluation

Our experiments show that ISSR outperforms baseline models. However, its ability to generate both plausible and valid distractors remains uncertain. To assess this, we tested ISSR-generated distractors on 30 GSAT exam questions. Thirteen Asia University students with strong GSAT English performance and sufficient proficiency were invited. To eliminate the possibility of practice effects, all participants were explicitly confirmed to have never taken the specific GSAT exams used in this evaluation. Questions were selected based on pass rates, sorted, divided into three equal groups, and 10 were randomly chosen from each. The groups were organized as follows:

First Group: This group included 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 are 7.15.

Second Group: The questions in this group had pass rates ranging from 54% to 65%, with a standard deviation of 3.27.

Third Group: This group covered questions with the highest pass rates, from a minimum of 65% to a maximum of 87%. The standard deviation for this group was 5.30.

This approach assessed ISSR-generated distractors across varying difficulty levels, evaluating their effectiveness in different contexts. Participants labeled distractors they confidently identified as plausible correct answers, which could invalidate the question.

Table 10 presents the students' accuracy. The "Challenging questions" column indicates the number of questions labeled as confusing due to the generated distractors. Over half of the students scored between 90 and 100, and 6 different questions effectively misled them. Two students scoring 60–70 did not label any distractors as confusing. This suggests that high-proficiency students are more sensitive to subtle semantic differences, making distractor selection more challenging for them. In contrast, lower-proficiency students may not fully analyze the distractor options and instead rely on guessing, leading to a reduced perception of difficulty. These findings support the effectiveness of ISSR in creating meaningful distractors that challenge advanced learners while maintaining accessibility for lower-proficiency students.

8. Limitations

In this study, we used GPT-3.5 and Llama3 for comparison. These models were selected based on availability at the time of experimentation, as conducting controlled evaluations requires significant time. Notably, the ISSR framework is model-agnostic and can be applied to other models, such as o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025). ISSR's core value lies in the selection mechanism and self-review process, which help maintain test quality even as models continue to evolve.

The GSAT dataset serves as a representative case of standardized vocabulary assessments, and ISSR is designed as a methodological framework rather than being restricted to a specific dataset. While different exams may have varying structures and objectives, ISSR's approach can be adapted to broader assessment contexts with appropriate adjustments. The quality of the candidate set may depend on the candidate generation mechanism.

The prompts were manually crafted to ensure optimal performance. However, ISSR is designed as a fully automated system, offering a flexible tool that enables educators to interact with the system using natural language inputs, thereby eliminating the need for specialized technical expertise. Future research could explore adaptive prompt tuning mechanisms that dynamically adjust based on user feedback, refining the system's ability to align with educator expectations.

9. Conclusion

Vocabulary acquisition is key to mastering a second language, and effective vocabulary tests help learners reinforce understanding and identify knowledge gaps. However, crafting suitable distractors is time-consuming and inconsistent. This paper presents the ISSR framework to assist teachers in generating high-quality distractors. By integrating predefined filtering rules and leveraging LLMs for selection rather than direct generation, ISSR improves stability and validity by mitigating issues such as semantic inconsistency, difficulty imbalance, and redundancy. Compared to existing methods, ISSR achieves substantial performance gains. While ISSR requires initial manual prompt design, educators can adjust the system using natural language inputs, making it more adaptable to different assessment needs. Its self-review mechanism ensures minimal human intervention post-generation, making it more scalable for large-scale assessments. However, ISSR requires more computing resources than similar methods due to its use of both LLM and BERT, and its self-review mechanism, which verifies distractors through binary-choice conversion, slows generation. Future work will focus on addressing these efficiency and resource constraints. Additionally, since we employed the existing models as distractor generator, 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. Developing an advanced method is left for future work.

Ethics Statement

This study involved student participants who voluntarily took part in the annotation task after being fully informed about its nature and requirements. Prior to participation, all students were briefed on the purpose of the study, the details of the annotation process, and the

fact that no additional financial compensation would be provided. They provided informed consent, acknowledging their understanding of the task and their voluntary participation.

Acknowledgements

This research was partially supported by National Science and Technology Council, Taiwan, under grant NSTC 114-2221-E-A49-057-MY3.

References

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Christopher JC Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81, 2010.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. CDGP: Automatic cloze distractor generation based on pre-trained language model. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5835–5840, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

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

Christiane Fellbaum. WordNet: An Electronic Lexical Database. *MIT Press*, 2:678–686, 1998.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

John Brian Heaton. *Writing English Language Tests*. Longman, 1988.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. Distractor generation for multiple choice questions using learning to rank. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors,

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 284–290, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

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. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

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

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, 13(1):15, 2018.

P. Viola and M. Jones. 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*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990517.

Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 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*, SIGIR '17, page 515–524, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale cloze test dataset created by teachers. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

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