Difficulty-Controllable Cloze Question Distractor Generation

Seokhoon Kang[1], Yejin Jeon[3,4†], Seonjeong Hwang[1], Gary Geunbae Lee[1,2]

[1]Graduate School of Artificial Intelligence, POSTECH, South Korea
[2]Department of Computer Science and Engineering, POSTECH, South Korea
[3]Mila Quebec AI Institute, Canada
[4]McGill University, Canada
{sh.kang, seonjeongh, gblee}@postech.ac.kr
yejin.jeon@mila.quebec

Abstract

Multiple-choice cloze questions are commonly used to assess linguistic proficiency and comprehension. However, generating high-quality distractors remains challenging, as existing methods often lack adaptability and control over difficulty levels, and the absence of difficulty-annotated datasets further hinders progress. To address these issues, we propose a novel framework for generating distractors with controllable difficulty by leveraging both data augmentation and a multitask learning strategy. First, to create a high-quality, difficulty-annotated dataset, we introduce a two-way distractor generation process to produce diverse and plausible distractors. These candidates are filtered and then categorized by difficulty using an ensemble QA system. Second, this newly created dataset is used to train a difficulty-controllable generation model via multitask learning. Experimental results demonstrate that our method generates high-quality distractors across difficulty levels and substantially outperforms GPT-4o in aligning distractor difficulty with human perception.

1 Introduction

The widespread adoption of e-learning platforms has transformed traditional methods used in education, enabling access to knowledge and breaking down geographical and temporal barriers (de Souza Rodrigues et al., 2021). Moreover, with the growing number of learners, scalable and effective assessment methods have become increasingly critical. As such, cloze questions are being widely used in language proficiency evaluation for their ability to assess diverse linguistic skills. Specifically, this particular assessment format involves the removal of specific words from a passage and requires learners to fill in contextually appropriate terms (Taylor, 1953). Among their variants, multiple-choice cloze questions are particularly favored for their ease of scoring and enhanced objectivity in large-scale testing.

A key challenge in constructing such questions lies in generating plausible distractors, or incorrect options (Haladyna, 2004) that challenge students to engage with the material and enhance reading comprehension by distinguishing the correct answer from similar but incorrect choices. Distractors that are neither too obvious nor overly misleading are not only essential for maintaining assessment validity, but also serve as a critical determinant of overall question difficulty (Rezigalla et al., 2024; Susanti et al., 2017, 2016). However, manually crafting high-quality distractors is time-consuming and resource-intensive, which makes it impractical for large-scale deployment.

Towards this, several studies have proposed automated distractor generation systems (Yeung et al., 2019; Ren and Q. Zhu, 2021; Chiang et al., 2022; Wang et al., 2023). While effective at replicating distractors in the training data (Ren and Q. Zhu, 2021; Chiang et al., 2022; Wang et al., 2023), these approaches struggle to generate distractors with varying difficulty levels, which limits their use in personalized learning environments. Although Yeung et al. (2019) explored difficulty-aware distractor generation, they rely on predefined candidate lists which limits diversity and difficulty range.

The challenge pertaining to difficulty control arises from multiple factors. First, difficulty is inherently subjective and lacks a universally accepted metric, which complicates its operationalization in automated systems (AlKhuzaey et al., 2024). Second, the limited number of distractors per question in the dataset restricts the variety of examples available for training, which hinders the model's ability to generalize across a spectrum of difficulty levels.

To address these challenges, we propose a framework for difficulty-controllable distractor generation that combines data augmentation and multitask

[†]This work was conducted while at POSTECH.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18774–18791

learning. Since our framework targets language proficiency assessment, we define difficulty based on the semantic plausibility of the distractor within the context, which is consistent with the existing benchmark (Xie et al., 2018). This approach allows for more objective control, independent of individual learner differences such as vocabulary size, thereby mitigating the subjectivity commonly associated with difficulty modeling. Additionally, since the most commonly used datasets for cloze questions focus on word-level completions, and modifying a single word to achieve a continuous spectrum of difficulty poses inherent challenges, we adopt a binary classification approach for measuring distractor difficulty.

To construct a difficulty-annotated dataset, we introduce a two-way distractor generation method that incorporates an information restriction strategy to produce diverse and contextually plausible distractors. Candidates are refined through a filtering stage to ensure semantic and grammatical quality, and then clustered by difficulty using an ensemble of QA models with score normalization for improved accuracy. We then train a generation model on the augmented dataset using a multitask learning objective, enabling it to not only produce distractors at specified difficulty levels but also to distinguish between correct answers, easy, and hard distractors.

Quantitative and qualitative evaluation results demonstrate the effectiveness of our approach. Our augmentation method expands the CLOTH dataset (Xie et al., 2018) to an average of 12 distractors per question, and is categorized by difficulty. We also observe that our high-attention information restriction strategy allows for flexible control over both distractor difficulty and semantic coverage. Our method significantly outperforms GPT-4o, with 73.25% and 64.23% difficulty accuracy in generating hard and easy distractors, respectively. Our model also significantly reduces the invalid distractor ratio, with no invalid distractors in easy distractors and 1.6% for hard distractors.

Our contributions are summarized as follows:

- We introduce a novel two-way data augmentation pipeline that enables flexible difficulty control via information restriction.
- We design a multitask learning strategy that allows the model to distinguish answers from distractors and assess their relative difficulty.
- Through comprehensive automatic and human evaluations, we demonstrate that our proposed method significantly outperforms GPT-4o in aligning distractor difficulty while maintaining a low invalid distractor ratio.
- We release both the augmented dataset and the trained model to support future research.[1]

2 Related Work

Early studies used linguistic heuristics and domain-specific resources for distractor generation (Pino and Eskénazi, 2009; dos Santos Correia et al., 2010). In particular, they leveraged thesauri, taxonomies, or predefined vocabularies to select distractors that were semantically related to the correct answer. However, the reliance on domain-specific resources limited their applicability across diverse subjects. To address the limitations of such heuristic methods, subsequent studies incorporated general-purpose knowledge bases. For instance, Ren and Q. Zhu (2021) proposed a framework that leverages knowledge bases like Probase (Wu et al., 2012) and WordNet (Fellbaum, 1998) to generate candidate distractors. Their approach used a learning-to-rank model to select distractors and demonstrated improved performance across multiple domains. Despite these advancements, the dependency on structured knowledge bases still constrains the model's adaptability in domains where such resources are sparse or outdated.

Recent literature has shifted towards utilizing transformer-based pretrained language models (PLMs) for distractor generation. For instance, Chiang et al. (2022) used PLMs to generate distractors, and Wang et al. (2023) formulated cloze distractor generation as a Text2Text task and used PLMs enhanced with pseudo Kullback-Leibler divergence to prevent generating distractors that were too similar to each other. Nevertheless, these methods still struggle with difficulty control, leading to limited adaptability across varying educational needs.

As such, difficulty-controllable question generation has emerged as a relatively new research domain. With the growing popularity of PLMs, recent works such as Uto et al. (2023), Tomikawa and Uto (2024), and Park et al. (2024) combine PLMs with Item Response Theory (IRT) to model and control question difficulty. However, the correlation between human and PLM-simulated IRT parameters remains unverified in the cloze domain, and validating this is challenging since reliable IRT

[1] https://github.com/ksh108405/DCDG

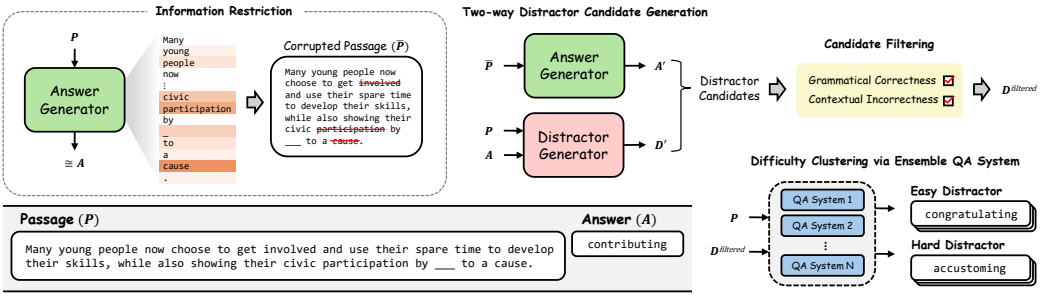

Figure 1: Overview of the dataset augmentation pipeline.

estimation necessitates large-scale human response data (Sireci, 1992), which is unavailable in public datasets. Furthermore, such methods incur high computational costs, often requiring hundreds of models to simulate student responses (Tomikawa and Uto, 2024). Given these constraints, we adopt a more resource-efficient approach using a discrete difficulty metric, which is calibrated with expert annotations.

3 Methodologies

3.1 Dataset Augmentation

To augment distractors for cloze questions, we propose a three-step system designed to 1) generate, 2) filter, and 3) cluster distractors by difficulty. Each step is detailed in its respective subsections. This augmentation process is illustrated in Figure 1.

Two-Way Candidate Generation In order to generate diverse and high-quality candidates for distractor augmentation, we adopt a two-way approach that integrates two distinct methods. First, we use a *distractor generator*, which is a language model fine-tuned on the distractors present in the original dataset. Yet, as this generator is specifically trained to replicate the characteristics of ground-truth distractors, its output is constrained by the semantic and difficulty distribution of the training data, which limits the generator's ability to produce diverse distractors for novel contexts or difficulty levels. Therefore, to address this limitation, we propose an information-restriction generation technique. This method re-purposes an *answer generator*, which was originally trained to generate correct answers, by strategically deleting its available information to produce distractors. This approach is built on two established findings; distractors deviate from the correct answer by contradicting specific parts of the passage (Ondov et al., 2024), and there exists a correlation in reading patterns between humans and transformer-based models (Zou et al., 2023; Bensemann et al., 2022).

Building on this insight, we repurposed the answer generator in a two-stage process to generate distractors. In the first stage, it processes the entire passage and generates the answer autoregressively. During this process, attention scores are computed at each generation step by summing across all layers and heads, and then accumulated across all generated answer tokens to derive the final attention scores for identifying crucial passage words. In the second stage, the passage is selectively pruned by removing words with the highest attention scores until a predefined deletion ratio is met. The pruned passage is then fed back into the answer generator to generate distractors that intentionally conflict with the passage. By adjusting the deletion ratio, we control the degree of inconsistency with the passage, thus producing distractors of varying difficulty. Pseudo-code for this information restriction is provided in Appendix B.

Candidate Filtering An essential property of effective distractors is that they must not align with the passage more closely than the correct answer. To achieve this, we use a two-step candidate filtering process that verifies whether each distractor is grammatically appropriate for the blank and ensures it cannot be a valid answer.

First, grammatically incorrect candidate distractors are eliminated using the rule-based LanguageTool (Naber, 2003) grammatical error corrector. Second, GPT-4o mini is used to assess the semantic plausibility of the remaining candidates by prompting the model to judge whether a distractor could serve as a valid answer in the given cloze passage. Those considered plausible answers are removed so as to retain only contextually inappropriate distractors. This two-step filtering process thus yields distractors that are both well-formed and clearly unsuitable as correct answers.

Difficulty Clustering To measure distractor difficulty, we use an ensemble of PLMs as done in prior studies (Yeung et al., 2019; Chiang et al., 2022; Çavuşoğlu et al., 2024). Each PLM in the system is fine-tuned with a multiple-choice cloze QA setup to assign a score to each option, which represents its likelihood of being selected as the correct answer. To cluster distractors by difficulty, we divide the entire pool into three equal-sized subsets based on their scores. The hard distractor set is comprised of distractors from the top third subset, while the easy distractor set is derived from the bottom third subset. To maximize diversity, we computed STS using cosine similarity on sentence embeddings from the `all-mpnet-base-v2` model in the Sentence-Transformer (Reimers and Gurevych, 2019). Specifically, we calculated a pairwise cosine similarity matrix for all distractor candidates and filtered out any option with a similarity score exceeding 0.8 with any other selected candidate. Distractors in the middle range are excluded from use, as they do not clearly align with either difficulty category.

A key aspect of the ensemble system is the normalization of confidence scores from each model, which is critical for ensuring balance across models. Given that the confidence scores for distractors exhibit a highly right-skewed distribution, we apply Box-Cox transformation (Box and Cox, 1964). This transformation is applied independently for each passage and model, with parameter λ optimized via maximum likelihood estimation to best fit a Gaussian distribution. For details, see Appendix E.

3.2 Training Strategy

To effectively train our difficulty-controllable distractor generation model using the augmented dataset and to enhance its understanding of distractor difficulty, we propose a novel multitask training strategy. This strategy consists of one main task and two auxiliary tasks, all formulated in a sequence-to-sequence (seq2seq) framework. Input-output formats for each task are shown in Figure 2.

Main Task At the core of our method is difficulty controllable distractor generation (DCDG), where an LLM is fine-tuned on the previously augmented dataset (§3.1). For each passage, the model is trained separately on the entire set of distractors from both the hard and easy subsets, treating each difficulty level as a distinct training instance.

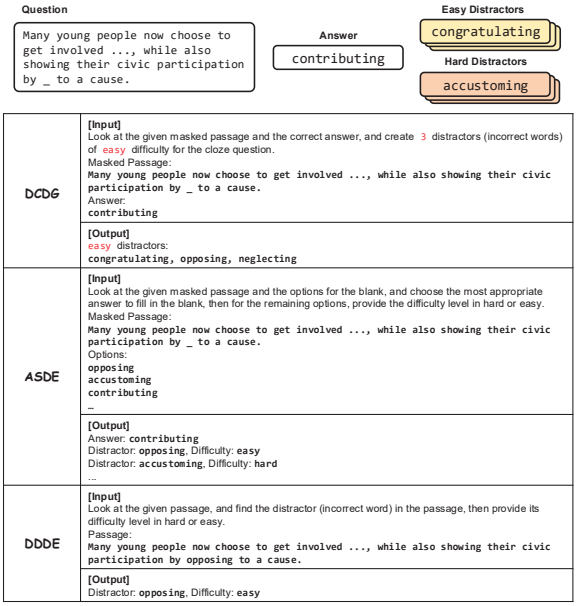

Figure 2: Overview of training methods of difficulty-controllable cloze distractor generation model. Template variables are highlighted in red.

Specifically, given the cloze passage, the number of distractors to generate, the desired difficulty level, and the ground-truth answer, the model learns to generate appropriate distractors that align with the specified difficulty.

Auxiliary Tasks To enhance the model's understanding of distractor semantics and difficulty, we introduce two auxiliary tasks inspired by Wang et al. (2023), and extend their formulation to explicitly incorporate difficulty signals. The primary goal of this multitask setup is to align the difficulty control tokens with the semantic plausibility of distractors within the given context, to differentiate between correct answers and distractors while recognizing relative difficulty. In the first auxiliary **answer selection and distractor difficulty estimation (ASDE)** task, the model learns to identify the correct answer and simultaneously assess the difficulty of given distractors. It receives a cloze passage with shuffled choices containing the ground-truth answer and distractors of varying difficulty. The model is trained to 1) identify the correct answer and 2) label each distractor with an estimated difficulty.

Unlike ASDE, in the **distractor detection and difficulty estimation (DDDE)** task, the model is given a cloze passage with a distractor word inserted into the blank. This task trains the model to 1) detect the distractor and 2) estimate its difficulty level, thereby aligning its predictions with the diffi-

culty annotated in the augmented dataset. Note that the full dataset is used to train the main task, while the auxiliary tasks are trained on their respective converted subsets. All tasks are optimized using a standard cross-entropy loss, calculated within the unified seq2seq formulation.

4 Experiments

Experiments are conducted on the CLOTH dataset (Xie et al., 2018), which is available for non-commercial research purposes only. This dataset consists of 7,131 unmasked passages and 99,433 cloze questions that are derived from middle and high school English exams in China, and covers a diverse range of topics and difficulty levels. Each passage contains multiple blanks for completion and is split into 5,513 passages for training, 805 for validation, and 813 for testing.

Two-Way Candidate Generation To augment the CLOTH dataset, we train Gemma 2 9B (Team, 2024) for both the distractor and answer generators. To prevent data leakage, we applied 5-fold cross-validation. For each iteration, the model was trained on four folds and performed inference exclusively on the remaining held-out fold. This cycle was repeated five times to augment the entire dataset without any overlap between training and generation data. To enhance the robustness of the answer generator during information restriction, 50% of the passages were randomly modified during training by removing words at varying rates between 0% and 100%. We used deletion ratios of [0.1, 0.2, 0.4] to create distractors of varying difficulty. Additionally, words surrounding the blank were preserved to avoid generating trivially easy distractors. Training was done with two NVIDIA A100-80GB GPUs, with a global batch size of 16 and a learning rate of 5e-6.

Candidate Filtering To ensure the quality of the augmented distractors, we applied a two-step filtering process. For grammatical validation, we used LanguageTool and deleted 2% of the candidates which were ungrammatical or nonsensical. To assess contextual correctness, we prompted GPT-4o mini in a 2-shot cloze QA format to detect distractors that fit the blank better than the original answer. This filtering was repeated three times, resulting in the removal of 21.31% of generated distractors. Prompt templates are detailed in Appendix D.

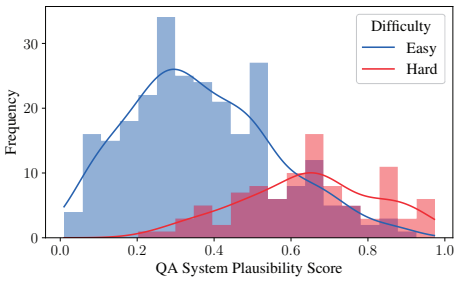

Figure 3: QA system annotation scores across two difficulty levels. The line represents the continuous distribution of the data using KDE.

Difficulty Clustering To determine the difficulty levels of the distractors in the augmented CLOTH dataset, we used a diverse set of small PLMs across 11 model families. A total of 18 models were trained and evaluated. To select the best-performing models which aligns with target users, a highly proficient ESL expert with over 20 years of academic English experience annotated the calibration set. Specifically, the expert labeled 359 distractors from 10 randomly sampled questions in the augmented validation set, categorizing them into two difficulty levels (easy and hard) based on perceived difficulty for ESL learners. Each model was then evaluated using this labeled test set. For the evaluation metric, we measure the degree of separation between the two difficulty distributions by computing the best balanced accuracy at the optimal threshold that maximally distinguishes them.

Among the top-performing model combinations, we selected those with a standard deviation below 0.5% for stability. This led to the selection of six models, including albert-xlarge-v2, albert-xxlarge-v2, conv-bert-base, electra-large-dis, roberta-large, and xlnet-large. Similarly, a 5-fold training approach was applied to prevent data leakage. For specific training details, refer to Appendix C.

Figure 3 presents the score distribution of the QA ensemble system for human-labeled distractors. The distribution of easy distractors is left-skewed relative to that of hard distractors, indicating that they tend to receive lower scores. This suggests that the system effectively identifies easy distractors as less plausible, demonstrating its ability to capture and reflect varying difficulty levels in alignment with human-labeled classifications.

Difficulty-controllable Distractor Generation Gemma 2 9B is trained on the augmented data for difficulty-controllable distractor generation using a

Dataset		Original	Augmented	
			Easy	Hard
# of Distractors per Question	μ	2.998	12.06	12.02
	σ	0.063	3.758	3.770
Similarity with Answer	μ	0.3504	0.2901	0.3592
	σ	0.1006	0.0600	0.0756

Table 1: Statistics of the augmented dataset. μ and σ denotes the mean and standard deviation, respectively.

Dataset		Original	Augmented	
			Easy	Hard
Relative Difficulty	Hardest	26.53%	3.42%	**70.05%**
	Middle	**52.56%**	23.32%	24.12%
	Easiest	21.21%	**73.17%**	5.63%
Invalid Ratio		0.9%	**0.0%**	4.2%

Table 2: Relative difficulty and invalid distractor ratios for ground-truth and augmented distractors.

Method	Answer Generator w/ IR	Distractor Generator
# of Distractors	19.25	29.66
Semantic Diversity	0.6928	0.6684
Semantic Overlap	0.2908	
Jaccard Overlap	0.1281	

Table 3: Comparison of distractor statistics between the answer generator with information restriction and distractor generator. Both Jaccard and Semantic Overlap quantify the relationship between the two generation methods.

single NVIDIA L40S GPU with a global batch size of 16. The learning rate was set to 5e-5 for DDDE multitask training and 3e-5 for the others. We also used early stopping to ensure proper evaluation across training objectives. To enhance efficiency, we applied low-rank adaptation (Hu et al., 2022) with $r = 16$ and $\alpha = 16$ and a 0.1 warm-up ratio for efficiency and stability.

5 Results and Analysis

5.1 Dataset Augmentation

Dataset Statistics Table 1 demonstrates the distinct improvements over the original CLOTH dataset in both distractor quantity and quality. The augmented dataset contains a greater number of distractors per question, and thus provides a more comprehensive set. Additionally, STS analysis with the `all-mpnet-base-v2` shows that easy distractors in the augmented dataset are less similar to the correct answers compared to those in the original dataset, whereas hard distractors exhibit greater similarity. Moreover, lower standard deviation of similarity scores across both difficulty levels suggests a more consistent control over the range of distractor difficulty.

Difficulty and Quality Evaluation We used GPT-4o (OpenAI, 2024) to further evaluate the quality and difficulty of augmented distractors. From the CLOTH test set, we sampled 1,000 questions and presented GPT-4o with a cloze passage alongside four options, which are the ground-truth answer, one randomly selected original distractor, and one easy and hard distractor from the augmented dataset. To assess relative difficulty, GPT-4o ranked the options based on their fit in the blank. Additionally, we measured the invalid ratio by checking whether any option was equally or more appropriate than the ground-truth answer. All evaluations were conducted in a 1-shot setting for consistency. The evaluation prompts are provided in Appendix G.

Table 2 reveals clear differences in difficulty across distractor types. Augmented hard distractors were categorized as the `Hardest` 70.05% of the time, compared to only 26.53% for original distractors. Similarly, easy distractors were classified as the `Easiest` in 73.17% of cases, in contrast to 21.21% in the original dataset. Even without explicit difficulty calibration, more than half of the original distractors fell between the augmented easy and hard ones. The invalid ratio for hard distractors was low at 4.2%, and no invalid options were found in easy distractors.

Analysis of Two-Way Candidate Generation As shown in Table 3, the statistics of the generated distractor candidates highlight the complementary strengths of our two-way candidate generation approach. The Jaccard overlap, which is calculated based on exact matches, is 12.81%. Moreover, the semantic overlap, which is measured by the average STS score between distractors for each question, is 0.2908. These relatively low values indicate that the answer generator with information restriction and the distractor generator produce distinct and minimally redundant distractor sets. Moreover, distractors that were generated from the answer generator show slightly greater semantic diversity compared to those from the distractor generator. This suggests that applying information restriction

Deletion Ratio	Diversity	Plausibility
0.1	0.6554	0.3404
0.2	0.6635	0.3189
0.3	0.6682	0.3066
0.4	0.6710	0.2978
0.5	0.6734	0.2920

Table 4: Average statistics for generated distractors under varying information deletion ratios.

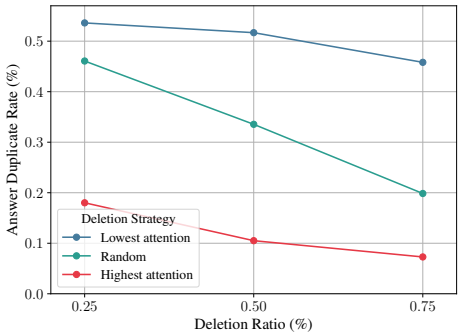

Figure 4: Duplication rate of generated distractors with the correct answers.

on the answer generator increases the semantic variety of the generated distractors and broadens difficulty levels. Further details on the difficulty distributions are provided in Appendix A.

Effect of Information Deletion Ratio To investigate the impact of information deletion ratios on the quality and diversity of distractors generated with information restriction, we experiment with five deletion ratios: [0.1, 0.2, 0.3, 0.4, 0.5]. For each set of distractors generated from an input question and answer, we evaluated two aspect of semantic diversity and plausibility. Semantic diversity is defined as one minus the average pairwise semantic similarity among the generated distractors, while plausibility is measured by the semantic similarity between a distractor and the ground truth answer.

Table 4 shows that increasing the information deletion ratio increases semantic diversity of distractors, while their plausibility with respect to the ground-truth answer decreases. This indicates that higher deletion ratios enhance distractor diversity, while lower ratios enhance their plausibility. These findings confirm that deletion ratio adjustments provide a flexible means of controlling distractor difficulty and semantic coverage.

Effectiveness of High-attention Deletion To further validate the effectiveness of our highest attention deletion strategy in augmenting distractors, we compared it with two alternative deletion approaches, which are random deletion and lowest attention deletion. As shown in Figure 4, deleting words with low attention scores or randomly selected words frequently led to high duplication rates of generated distractors matching the ground truth answer, even at relatively high deletion ratios. Specifically, these strategies resulted in more than 40% duplication rates when 25% of words were deleted. In contrast, our proposed high-attention deletion approach has significantly low duplication rates, which maintains an answer duplication rate below 20% at the same deletion ratio. This indicates that selectively deleting high-attention words effectively prevents the answer generator from reproducing the answer, thereby enhancing the efficiency of augmenting plausible distractors.

5.2 Difficulty-controllable Distractor Generation

Automatic Evaluation To evaluate the difficulty of the generated distractors, we used GPT-4o to assess their relative difficulty. Following the same methodology as the augmented dataset, we presented GPT-4o with cloze passages containing four options: the ground-truth answer, one original distractor, one easy distractor, and one hard distractor generated by our model. Additionally, we compared with those produced by GPT-4o in both 0-shot and 5-shot settings. In the 5-shot setup, we provided examples from the augmented dataset, including eight easy and eight hard distractors for five randomly selected questions. The detailed prompt and few-shot setup used for GPT-4o distractor generation are provided in Appendix F.

Table 5 highlights the effectiveness of multitask training in generating distractors with distinct difficulty levels. The model trained with multitask learning outperforms the one fine-tuned only on the main task, showing a greater ability to differentiate between the hardest and easiest distractor categories. In the hard-difficulty setting, 73.25% of the distractors from the ASDE + DDDE multitask model classify as `Hardest`, an improvement from 70.35% in the single task setup. Similarly, for easy distractors, the proportion categorized as `Easiest` increases from 58.49% to 64.23%. On the other hand, GPT-4o struggles with consistency, with only 33.54% (0-shot) and 46.39% (5-shot) of easy distractors classified as `Easiest` For hard distractors, GPT-4o's `Hardest` rate remains at 56.77% (0-shot) and 53.81% (5-shot), which is significantly lower than the proposed approach.

Method	Type	Relative Difficulty		
		Hardest	Middle	Easiest
GPT-4o (0-shot)	Original	13.64%	35.56%	51.31%
	Easy	29.60%	36.36%	33.54%
	Hard	56.77%	28.08%	15.15%
GPT-4o (5-shot)	Original	24.35%	38.38%	37.78%
	Easy	21.84%	31.06%	46.39%
	Hard	53.81%	30.56%	15.83%
DCDG	Original	20.40%	47.94%	32.36%
	Easy	9.25%	32.06%	58.49%
	Hard	70.35%	20.00%	9.15%
DCDG + ASDE	Original	19.64%	49.70%	31.66%
	Easy	8.22%	30.56%	60.72%
	Hard	72.14%	19.74%	7.62%
DCDG + DDDE	Original	22.69%	43.98%	33.73%
	Easy	9.34%	32.63%	57.53%
	Hard	67.97%	23.39%	8.73%
DCDG + ASDE, DDDE	Original	20.04%	**51.80%**	28.46%
	Easy	6.71%	28.96%	**64.23%**
	Hard	**73.25%**	19.24%	7.31%

Table 5: Relative difficulty evaluation results of generated distractors. *Original* denotes the distractors in the original CLOTH dataset.

Method	Invalid Ratio	
	Easy	Hard
GPT-4o (0-shot)	6.8%	16.9%
GPT-4o (5-shot)	1.6%	6.8%
DCDG	0.9%	6.4%
DCDG with ASDE	**0.1%**	7.0%
DCDG with DDDE	0.5%	6.3%
DCDG with ASDE + DDDE	0.2%	**5.1%**

Table 6: Invalid distractor evaluation results of generated distractors.

As shown in Table 6, the proposed model with multitask training also achieves notably low invalid ratio of only 0.2% for easy distractors and 5.1% for hard distractors. This significantly outperforms GPT-4o, where invalid distractors reach 16.9% (0-shot hard) and 6.8% (5-shot hard).

Furthermore, we evaluated the quality of generated distractors by qualitative study and compared our method against established baselines on the original dataset, which confirmed their high quality and consistency. Finally, to further examine the generalizability of our method, we conducted additional experiments using other small LLMs (sLLMs), which showed strong performance in generating high-quality distractors at controlled difficulty levels. For details, see Appendix L, H, and I respectively.

Source	Type	Relative Difficulty			Invalid Ratio
		Hardest	Middle	Easiest	
Dataset	Original	28.8%	**61.6%**	9.6%	3.2%
	Easy	16.8%	1.6%	**81.6%**	**0.0%**
	Hard	**54.4%**	36.8%	8.8%	1.6%
Model	Original	30.4%	**51.2%**	18.4%	1.6%
	Easy	24.0%	3.2%	**72.8%**	**0.0%**
	Hard	**45.6%**	45.6%	8.8%	1.6%

Table 7: Human evaluation results of both our augmented dataset and trained model.

Option Type	Chosen Ratio
Ground-Truth Answer	89.6%
Hard Distractor (Ours)	8.6%
Ground-Truth Distractor	1.6%
Easy Answer (Ours)	0.2%

Table 8: Mean ratio of being chosen as the answer for each option type.

Human Evaluation To assess the student-perceived cognitive load of distractors, we recruited five English as a Second Language (ESL) adult learners. Each participant had less than five years of experience using English. Additional details of human annotation setup is detailed in Appendix J.

For evaluation, 50 questions were randomly sampled from the CLOTH test set. Among these, 25 cloze questions were used to evaluate the distractors generated in the augmented dataset, while the remaining 25 questions assessed distractors produced by the multitask-trained model using DCDG with both ASDE and DDDE objectives. Participants were instructed to solve cloze questions before proceeding with the evaluation. Following the methodology used in our automatic evaluation with GPT-4o, the participants were asked to compare three distractors, which are the ground-truth distractor, an easy distractor, and a hard distractor. The evaluation consisted of two tasks:

1. **Relative Difficulty Assessment:** For each distractor, participants ranked its difficulty as "`Hardest`," "`Middle`," or "`Easiest`".

2. **Validity Check:** Participants indicated whether any distractors were more suitable for the blank than the ground-truth answer.

Table 7 and 8 presents the results of human evaluation on distractor difficulty and validity. These results align with GPT-4o assessments, confirming

that distractors labeled Hard and Easy were consistently perceived as the most and least challenging, respectively. Furthermore, in terms of distractor efficacy, ESL learners selected our Hard distractors more often than ground-truth and Easy distractors. This trend is notably more pronounced compared to native experts (Table 15), validating the model's effectiveness for the target ESL audience. Additionally, the invalid distractor ratio remained at or below 1.6%. The results demonstrate that our augmentation method maintains and improves the validity of the original CLOTH dataset.

In addition to evaluations with ESL users, we also conducted a complementary study involving ten proficient English speakers, which confirms pedagogical validity and the consistency of difficulty control across varying user proficiency levels. For details, see Appendix K.

Agreement Between Automatic and Human Evaluation To quantify the agreement between automatic and human evaluation, we used the same 50 questions from the human evaluation. For each question, the difficulty rankings from the five ESL annotators were averaged to produce a mean rank per distractor, and Spearman's rank correlation coefficient was computed against GPT-4o's rankings across all questions. The resulting GPT-Human correlation of 0.54 is comparable to the inter-human agreement of 0.62. Given that this task involves ranking only three distractors per question, where even a single-position swap substantially affects the rank correlation, the close alignment between GPT-Human and Human-Human agreement demonstrates that GPT-4o serves as a reliable proxy for human evaluation in this context.

6 Conclusion

In this paper, we have proposed a novel method for generating difficulty-controllable distractors, and addressed key challenges in automated assessment. Our data augmentation pipeline improved distractor plausibility and diversity through a two-way candidate generation strategy, and ensured quality via a filtering process. By incorporating an advanced normalization method into the difficulty clustering step, we were able to construct a cloze distractor dataset that is annotated with varying difficulty levels. We further designed a multitask learning framework that enabled the model to generate distractors with controlled difficulty. The resulting model outperformed GPT-4o in both automated and human evaluations, which produced valid distractors that align closely with human-perceived difficulty. This work contributes to scalable and adaptive assessment in e-learning, supporting systems that tailor distractor difficulty to learner profiles and domain-specific needs.

7 Limitations

Consideration of Passage Structure on Difficulty While item difficulty is influenced by a wide range of factors, including passage syntax and blank position, this study focuses exclusively on distractor difficulty. We prioritized this dimension as it is a critical factor in item development that is less dependent on learner variability. Future work could enhance the system by incorporating passage readability metrics and syntactic-semantic analyses of blank positions to achieve a more comprehensive difficulty control.

Generalization to Other Question Types The information restriction technique is designed for cloze-style questions to assess language proficiency, and its effectiveness for other formats, such as open-ended or math domain questions, are areas that require further study. Similarly, the deletion ratio approach may require additional adaptation to ensure meaningful distractor generation across different structures. Future work could refine deletion strategies or incorporate linguistic constraints to enhance generalizability.

Control Over Continuous Difficulty Levels While continuous metrics offer granularity, mapping them directly to pedagogically appropriate levels poses another challenges. Given the lack of prior research and benchmarks for difficulty control, and to avoid arbitrary thresholding, we adopted a binary classification approach grounded in expert calibration. Future work could extend this by leveraging normalized PLM ensemble scores from our difficulty clustering module to define finer-grained, pedagogically calibrated levels.

8 Ethical Considerations

Our system is designed as a teacher-centric authoring tool rather than a fully automated test generator. Teachers are responsible for reviewing, revising, and approving all content prior to use. The purpose of the system is not to replace human judgment, but to reduce the time and effort required for educators to create high-quality distractors.

Acknowledgments

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of an AI-Based Korean Diagnostic System for Efficient Korean Speaking Learning by Foreigners, Project Number: RS-2025-02413038, Contribution Rate: 45%); by the IITP (Institute of Information & Coummunications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00437866, Contribution Rate: 45%); and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH), Contribution Rate: 10%).

References

Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye Gaze and Self-attention: How Humans and Transformers Attend Words in Sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

G. E. P. Box and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Devrim Çavuşoğlu, Seçil Şen, and Ulaş Sert. 2024. DisGeM: Distractor Generation for Multiple Choice Questions with Span Masking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9714–9732, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 5835–5840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Aurélio de Souza Rodrigues, Paula Chimenti, and Antonio Roberto Ramos Nogueira. 2021. An Exploration of eLearning Adoption in the Educational Ecosystem. *Education and Information Technologies*, 26(1):585–615.

Rui Pedro dos Santos Correia, Jorge Baptista, Nuno Mamede, Isabel Trancoso, and Maxine Eskenazi. 2010. Automatic Generation of Cloze Question Distractors. In *Second Language Studies: Acquisition, Learning, Education and Technology (L2WS 2010)*, pages paper P2–11.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

T.M. Haladyna. 2004. *Developing and Validating Multiple-choice Test Items*. Lawrence Erlbaum Associates.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

D. Naber. 2003. *A Rule-Based Style and Grammar Checker*. GRIN Verlag.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2024. Pedagogically Aligned Objectives Create Reliable Automatic Cloze Tests. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3961–3972, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics.

Juan Pino and Maxine Eskénazi. 2009. Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1. In *ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2009, Warwickshire, England, UK, September 3-5, 2009*, pages 65–68. ISCA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4339–4347.

Assad Ali Rezigalla, Ali Mohammed Elhassan Seid Ahmed Eleragi, Amar Babikir Elhussein, Jaber Alfaifi, Mushabab A. ALGhamdi, Ahmed Y. Al Ameer, Amar Ibrahim Omer Yahia, Osama A. Mohammed, and Masoud Ishag Elkhalifa Adam. 2024. Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items. *BMC Medical Education*, 24(1):445.

Stephen G Sireci. 1992. The Utility of IRT in Small-Sample Testing Applications. In *The Annual Meeting of the American Psychological Association 100th*. ERIC.

Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Obari Hiroyuki. 2016. Item Difficulty Analysis of English Vocabulary Questions. In *Proceedings of the 8th International Conference on Computer Supported Education - Volume 1: CSEDU*, pages 267–274. INSTICC, SciTePress.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Controlling Item Difficulty for Automatic Vocabulary Question Generation. *Research and Practice in Technology Enhanced Learning*, 12(1):25.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

Yuto Tomikawa and Masaki Uto. 2024. Difficulty-Controllable Multiple-Choice Question Generation for Reading Comprehension Using Item Response Theory. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 312–320, Cham. Springer Nature Switzerland.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor Generation based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491, Toronto, Canada. Association for Computational Linguistics.

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*, SIGMOD '12, page 481–492, New York, NY, USA. Association for Computing Machinery.

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*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Chak Yan Yeung, John Lee, and Benjamin Tsou. 2019. Difficulty-aware Distractor Generation for Gap-Fill Items. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164, Sydney, Australia. Australasian Language Technology Association.

Jiajie Zou, Yuran Zhang, Jialu Li, Xing Tian, and Nai Ding. 2023. Human Attention During Goal-Directed Reading Comprehension Relies on Task Optimization. *eLife*, 12:RP87197.

A Differences Between Two Generation Methods

This section focuses on the difficulty range of distractors generated by two distinct methods. We assess the difficulty by calculating the QA ensemble system scores for each method. To visualize the difficulty distribution differences, we present a histogram of the score differences between the two generation methods.

Figure 5 illustrates that the answer generator yields more distractors at both difficulty extremes, indicating a broader difficulty spectrum. This finding suggests that restricting the available information allows the answer generator to generate both easier and more challenging distractors compared to the distractor generator, resulting in a more diverse and balanced set of options. This diversity enhances the robustness of the augmented dataset for training difficulty-controllable distractor generation models.

B Pseudo Code of Information Restriction Generation

Algorithm 1 presents the procedure for generating difficulty-controlled distractors by selectively restricting key information in a given passage. The method begins by identifying important words based on attention scores S assigned by the answer generator $\mathcal{M}_A$. The algorithm ranks the passage words P in descending order of significance according to these attention scores. By iteratively removing a subset of the most influential words at

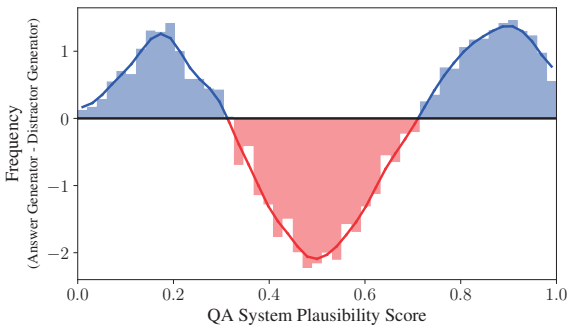

Figure 5: Histogram difference of QA ensemble system scores between two generation methods. Positive values (blue) indicate higher frequencies for the answer generator, while negative values (red) indicate higher frequencies for the distractor generator.

Algorithm 1 Information Restriction Generation

Input: Passage Words P, Answer Generator $\mathcal{M}_A$, Deletion Ratios R

Output: Distractors $\mathcal{D}$

Step 1: Sort P with attention scores S
$A, S \leftarrow \mathcal{M}_A(P)$
$P_{sorted} \leftarrow sort(P, S, descending)$
$\mathcal{D} \leftarrow \emptyset$
for $r \in R$ **do**
  Step 2: Remove key words from passage
  $n \leftarrow \lfloor r \cdot |P| \rfloor$
  $P_r \leftarrow P$
  for $i = 1$ to n **do**
    $P_r \leftarrow P_r - P_{sorted}[i]$
  end for
  Step 3: Generate distractors
  $d_r \leftarrow \mathcal{M}_A(P_r)$
  $\mathcal{D} \leftarrow \mathcal{D} \cup d_r$
end for
Return: $\mathcal{D}$

varying deletion ratios $r \in R$, the algorithm creates pruned passages P_r with different levels of missing information. This ensures that the generated distractors are influenced by controlled deletions in contextual cues.

Once the key words are removed, the pruned passage P_r is reprocessed by the answer generator to generate plausible distractors d_r. By varying the deletion ratio r, the algorithm produces a diverse set of distractors with different levels of semantic plausibility. This controlled variation significantly enhances their suitability for difficulty-tunable multiple-choice questions.

Model Name	Params	LR	Eval Loss
albert-xlarge-v2	58M	3e-06	0.6943
albert-xxlarge-v2	223M	5e-06	0.4712
bert-base-uncased	110M	3e-05	1.3764
bert-large-uncased	336M	1e-05	1.1166
conv-bert-base	106M	3e-05	1.2404
deberta-v2-xlarge	900M	1e-06	0.5639
deberta-v2-xxlarge	1.5B	1e-06	0.4986
distilbert-base	67M	3e-05	1.8335
distilroberta-base	83M	3e-05	1.5995
electra-base-discriminator	110M	1e-05	1.3611
electra-large-discriminator	336M	1e-05	0.6205
mpnet-base	133M	1e-05	1.0518
roberta-base	125M	7e-06	1.1444
roberta-large	355M	3e-06	0.9509
spanbert-base-cased	110M	1e-05	1.4093
spanbert-large-cased	336M	1e-05	0.9545
xlnet-base	110M	1e-05	1.1867
xlnet-large	336M	1e-05	0.7915

Table 9: Overview of the model candidates used in the ensemble QA system, including the model names, parameter sizes, tuned learning rates (LR), and evaluation loss on the validation set.

C Detailed Information of Ensemble QA System

To construct a robust and accurate ensemble QA system for estimating distractor difficulty, we evaluated 18 encoder-only models spanning 11 model families with parameter sizes ranging from 67M to 1.5B. Each model was fine-tuned using a batch size of 8 and early stopping, alongside an optimized learning rate to enhance performance and prevent overfitting. ALBERT and DeBERTa models were trained on an NVIDIA RTX6000ADA GPU, while all other models trained on an NVIDIA L40S GPU. Table 9 presents a detailed overview of the models, including their respective tuned learning rates and evaluation loss scores measured on a validation set.

D Prompts for Candidate Filtering

The candidate filtering process use a two-shot prompting approach with GPT-4o mini to refine the set of distractor candidates by eliminating those that were more suitable than the ground-truth answer. The detailed prompt structure used for this filtering process is presented in Figure 6. This prompt instructed the model to evaluate each candidate independently, provide a justification for its appropriateness, and identify a subset of acceptable candidates. Finally, any distractors included in the list were removed from the candidate pool.

```
[SYSTEM]
Your task is to find every possible answer for the cloze question from the candidates
    provided.
For each candidate, provide a brief explanation about the appropriateness of the candidate.
Then, at the end, **must** provide "Appropriate candidates", with every candidate that could
    be considered as correct, seperated by a new line.
If there is no appropriate candidate, provide "Appropriate candidates: None".
Evaluate each candidate independently.

[USER]
Masked passage:
<passage with a blank indicated as _____>

Candidates:
<a ground-truth answer and distractor candidates, separated by line break>
```

Figure 6: Prompt template used for filtering distractor candidates with GPT-4o mini.

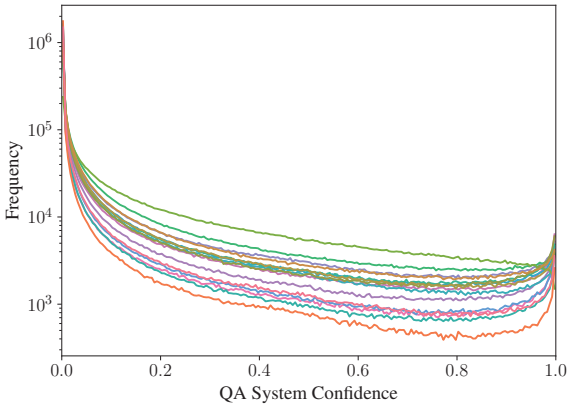

Figure 7: Raw confidence score distribution for each PLM in a log scale.

E Normalization in QA Ensemble System

Normalization of confidence scores across multiple PLMs is crucial to mitigate scale imbalances arising from model-specific and question-dependent variations. Variations in model performance often yield disparate confidence distributions, especially across differing model capacities and question difficulties. To address this, we analyzed the confidence distribution of distractors, which revealed a pronounced right skew. We then compared two normalization techniques, Box-Cox transformation and ranking normalization, to evaluate their effectiveness in stabilizing score distributions.

The confidence score distributions of distractors across PLMs exhibit significant right skewness, as shown in Figure 7, with the y-axis on a log scale. The skewness and variations in model-specific distributions indicate inconsistencies in how PLMs assess distractor plausibility. Such discrepancies can reduce the effectiveness of an ensemble system by introducing aggregation biases.

To address this issue, two normalization methods are considered. Ranking normalization replaces confidence scores with their ranks, effectively mitigating scale imbalance in a simple strategy. However, it only preserves ordinal relationships and disregards absolute score differences. In contrast, Box-Cox normalization transforms right-skewed distributions into a Gaussian form, allowing the retention of relative score magnitudes. To evaluate their effectiveness, we compare the two methods based on their ability to achieve optimal separation in confidence distributions.

The results in Table 10 indicate that Box-Cox normalization achieves a higher degree of separation compared to ranking normalization. Moreover, model combinations with a standard deviation below 0.5% consistently exhibit higher separation scores under Box-Cox normalization, demonstrating its stability across multiple questions. These findings suggest that Box-Cox normalization more effectively balances score scales, improving the robustness of confidence-based ranking in the QA ensemble system.

F GPT-4o Distractor Generation Prompt

To evaluate difficulty-controllable distractor generation with GPT-4o, we adopted a 5-shot prompting strategy. For each of the five few-shot examples, we provided both 'Easy' and 'Hard' distractor sets, resulting in a total of ten difficulty-labeled example instances. This design allows the model to learn specific difficulty boundaries by directly contrasting the Easy and Hard examples within the context.

The system prompt which used for GPT-4o distractor generation is in Figure 8.

Box-Cox Normalization			Ranking Normalization		
Model List	μ	σ	**Model List**	μ	σ
albert-xxlarge-v2, bert-large, roberta-large, xlnet-large	0.841	0.007	albert-xxlarge-v2, bert-large, roberta-large, xlnet-large	0.836	0.007
albert-xxlarge-v2, conv-bert-base, roberta-large	0.840	0.006	albert-xxlarge-v2, deberta-v2-xxlarge, mpnet, roberta-large, spanbert-base, xlnet-large	0.836	0.006
albert-xxlarge-v2, xlnet-large	0.840	0.008	albert-xxlarge-v2, conv-bert-base, roberta-large, xlnet-large	0.836	0.007
albert-xxlarge-v2, roberta-large, spanbert-base, xlnet-large	0.840	0.007	albert-xxlarge-v2, conv-bert-base, deberta-v2-xlarge, roberta-large, spanbert-base	0.836	0.005
albert-xxlarge-v2, roberta-large	0.839	0.007	albert-xxlarge-v2, conv-bert-base, roberta-large	0.836	0.006
albert-xxlarge-v2, conv-bert-base, roberta-large, xlnet-large	0.838	0.008	albert-xlarge-v2, albert-xxlarge-v2, mpnet, roberta-large, spanbert-base, xlnet-large	0.835	0.009
albert-xlarge-v2, albert-xxlarge-v2, conv-bert-base, electra-large-dis, roberta-large, xlnet-large	**0.837**	**0.004**	albert-xlarge-v2, albert-xxlarge-v2, bert-large, roberta-large, spanbert-base, xlnet-large	0.834	0.012
albert-xlarge-v2, albert-xxlarge-v2, bert-large, conv-bert-base, deberta-v2-xxlarge, electra-large-dis, roberta-large, spanbert-base	0.836	0.004	albert-xlarge-v2, albert-xxlarge-v2, spanbert-base, xlnet-large	0.834	0.007
albert-xlarge-v2, albert-xxlarge-v2, bert-large, conv-bert-base, deberta-v2-xlarge, electra-large-dis, roberta-large, spanbert-base	0.836	0.009	albert-xlarge-v2, albert-xxlarge-v2, conv-bert-base, deberta-v2-xlarge, roberta-large, spanbert-base, xlnet-large	**0.834**	**0.004**
albert-xxlarge-v2, conv-bert-base, roberta-large, spanbert-base	0.836	0.005	albert-xxlarge-v2, conv-bert-base, deberta-v2-xxlarge, electra-large-dis, roberta-base, roberta-large, spanbert-base, xlnet-large	0.834	0.004

Table 10: Comparison of the top 10 model combinations by degree of separation under different normalization methods. In the table μ presents the mean and σ presents the standard deviation of performance. For each normalization method, the best-performing model combination with $\sigma < 0.5\%$ is highlighted in bold.

```
[SYSTEM]
Look at the given masked passage and the correct answer, then create 8 distractors (incorrect
    words) of given difficulty ("Easy" or "Hard") for the cloze question.
At the end, you must write "Easy Distractors:" or "Hard Distractors:", then write down the
    all the distractors you made, seperated by a new line.

[USER]
Masked passage:
<passage with a blank indicated as _____>

Answer:
<ground-truth answer>

Difficulty:
<'Easy' or 'Hard'>
```

Figure 8: Prompt template used for GPT-4o distractor generation with difficulty control.

G GPT-4o Evaluation Prompts

To ensure the effective evaluation of cloze question distractors, we developed two structured prompt templates designed for GPT-4o. The prompt in Figure 9 assesses the relative difficulty of distractors by ranking them according to how easily they might be confused with the ground truth answer, by providing explanations for why each distractor is incorrect. The second prompt in Figure 10 detects distractors that are equally or more appropriate than the ground-truth answer, thereby filtering out invalid options that might compromise the overall test's validity.

H Consistency with Original Dataset

Since standard multiple-choice questions typically provide only a limited number of distractors per question, categorizing them into difficulty levels results in severe data sparsity. Therefore, while data augmentation is essential to overcome this scarcity, it is still necessary to verify that the generated distractors remain highly consistent with the original dataset. To address this, we evaluated the alignment of the model with both the original and the augmented CLOTH datasets by measuring its ability to generate distractors that replicate the distractor in each dataset.

```
[SYSTEM]
Given a cloze question consisting of masked passage, four options and correct answer,
    evaluate the incorrect options by their **relative** difficulty.
For each incorrect options, provide a brief explanation about their incorrectness.
Then, at the end, you must write "Results:", then write down the all incorrect options in the
    order of **relative** difficulty.
Start by writing the incorrect answer that is most confusing to distinguish from the correct
    answer.

[USER]
Masked passage:
<passage with a blank indicated as _____>

Options:
<four options, separated by line break>

The answer is: <ground-truth answer>
```

Figure 9: Evaluation prompt template for relative difficulty.

```
[SYSTEM]
Given a cloze question consisting of masked passage, three options and correct answer,
    evaluate the option if they are equally suitable or just suitable as the correct answer.
For each options, provide a brief explanation about their incorrectness.
Then, at the end, you must write "Results:", then write down the all options that equally
    suitable or just suitable as the answer, seperated by a new line.
If there is no options that are equally suitable or just suitable as the answer, provide
    "Results: None".

[USER]
Masked passage:
<passage with a blank indicated as _____>

Options:
<three distractors, separated by line break>

Answer: <ground-truth answer>
```

Figure 10: Evaluation prompt template for assessing invalid distractors.

Specifically, we generated distractors for the test set using the trained model and measured performance via F1 score under an exact match criterion with both original and augmented distractors. Additionally, we compared the model's performance on the original dataset with the state-of-the-art methods (Chiang et al., 2022; Wang et al., 2023) to examine the impact of training on the augmented dataset. For the augmented test set, we further analyzed performance across easy and hard difficulty levels to assess the model's ability to capture difficulty-specific characteristics.

Table 11 shows that our models exhibit a slight decrease in F1 score compared to state-of-the-art on the original dataset, which is reasonable given that our model was trained on more diverse data. On the augmented dataset, our model significantly performs better, achieving F1 scores of 26.93 for easy distractors and 41.61 for hard distractors under the ASDE + DDDE multitask setup. The lower performance on easy distractors likely reflects their broader semantic variability. These results show that the model leverages augmentation to generate controlled distractors while avoiding answers that are misaligned with the original dataset.

I Experiments with Various sLLMs

To demonstrate the generalizability of our approach, we have conducted additional experiments. We fine-tuned two other sLLMs, Llama 3.1 8B and Qwen 2.5 7B, on our augmented dataset using the DCDG with ASDE + DDDE multitask training strategy. We then evaluated the relative difficulty and invalid ratio of the generated distractors using

Eval Dataset	Diff	Method	F1@10
Original	-	CDGP (2022)	**15.37**
		Text2Text (2023)	14.05
		DCDG	12.84
		DCDG with ASDE	12.88
		DCDG with DDDE	13.23
		DCDG with ASDE + DDDE	13.05
Augmented	Easy	DCDG	26.42
		DCDG with ASDE	26.56
		DCDG with DDDE	25.33
		DCDG with ASDE + DDDE	**26.64**
	Hard	DCDG	41.76
		DCDG with ASDE	41.73
		DCDG with DDDE	41.05
		DCDG with ASDE + DDDE	**41.98**

Table 11: Distractor exact match results on test set of original and augmented dataset. Diff denotes difficulty.

GPT-4o under the same experimental conditions.

As Table 12 and 13 show, when trained with our proposed method, both Llama 3.1 8B and Qwen 2.5 7B demonstrate a strong capability to generate distractors aligned with specified difficulty levels, while maintaining a low invalid ratio. Although all models perform well, Qwen 2.5 7B shows a slightly better performance than Gemma 2 9B, which confirms the effectiveness and model-agnostic potential of our framework.

J Details of Human Annotation

For all human evaluations conducted in this study, we recruited participants through the Amazon Mechanical Turk platform. To ensure ethical standards, participants were explicitly informed that the collected data would be used exclusively for research purposes before beginning the tasks. We compensated participants with a fixed hourly rate in accordance with the legal minimum wage standards of the authors' nationality.

Participants were categorized into two groups based on their English proficiency duration:

- **Experts:** Users with over 10 years of English usage experience.
- **ESL Learners:** Users with less than 5 years of English usage experience.

The demographic breakdown of the recruited participants is detailed in Table 14.

Method	Type	Relative Difficulty		
		Hardest	Middle	Easiest
GPT-4o (0-shot)	Original	13.64%	35.56%	51.31%
	Easy	29.60%	36.36%	33.54%
	Hard	56.77%	28.08%	15.15%
GPT-4o (5-shot)	Original	24.35%	38.38%	37.78%
	Easy	21.84%	31.06%	46.39%
	Hard	53.81%	30.56%	15.83%
Gemma 2 9B	Original	20.04%	**51.80%**	28.46%
	Easy	6.71%	28.96%	64.23%
	Hard	73.25%	19.24%	7.31%
Llama 3.1 8B	Original	22.24%	50.40%	27.76%
	Easy	6.81%	27.35%	**65.53%**
	Hard	70.94%	22.24%	6.71%
Qwen 2.5 7B	Original	19.64%	51.10%	29.46%
	Easy	6.81%	27.56%	65.33%
	Hard	**73.55%**	21.34%	5.21%

Table 12: Relative difficulty evaluation results from various sLLMs. In the table, *Original* denotes the distractors in the original CLOTH dataset.

Method	Invalid Ratio	
	Easy	Hard
GPT-4o (0-shot)	6.8%	16.9%
GPT-4o (5-shot)	1.6%	6.8%
Gemma 2 9B	0.2%	5.1%
Llama 3.1 8B	0.2%	5.2%
Qwen 2.5 7B	**0.0%**	**4.2%**

Table 13: Invalid distractor evaluation results from various sLLMs.

K Evaluation with English Expert

In this section, we present our experiment results from **expert** English users. We had recruited 10 proficient English users following Appendix J and had them evaluate 50 questions by choosing the most suitable answer from four options, which are the ground-truth answer, the ground-truth distractor, our easy distractor, and our hard distractor.

Table 15 and 16 demonstrate that our model's hard distractors are sophisticated enough to be chosen over the original dataset's distractors even by proficient users. The near-zero selection rate for easy distractors confirms that our difficulty control is effective across different user proficiency levels. Additionally, while our hard distractors are challenging than the original distractors, their invalid ratio remains low.

Demographic	Experts (N=10)	ESL (N=5)
Region		
North America	3	3
South America	5	2
Asia	2	-
Gender		
Male	8	4
Female	2	1

Table 14: Demographic statistics of human annotators.

Option Type	Chosen Ratio
Ground-Truth Answer	83.2%
Hard Distractor (Ours)	9.4%
Ground-Truth Distractor	7.0%
Easy Answer (Ours)	0.4%

Table 15: Mean ratio of being chosen as the answer by expert English user.

Option Type	Invalid Ratio
Hard Distractor (Ours)	8.4%
Ground-Truth Distractor	5.0%
Easy Answer (Ours)	1.0%

Table 16: Invalid ratio for each distractor type by expert English user.

L Qualitative Analysis

In this section, we present the results of two selected questions from the CLOTH test set. For each question, we provide the passage, the ground-truth answer, three original distractors, eight augmented distractors, and eight generated distractors on both easy and hard difficulty levels.

Table 17 demonstrates an example from the CLOTH-M subset, which is derived from middle school English examinations. It can be observed that both the augmented distractors and generated distractors contain a diverse range of distractors across easy and hard difficulty levels. The augmented easy distractors primarily consist of negative words that contrast with the passage's main theme of diverse travel experiences. In contrast, the augmented hard distractors include words that are more consistent with the passage's content. The model-generated easy distractors show a diverse semantic range but still include words that do not align well with the passage. Meanwhile, the hard distractors by the model exhibit a stronger semantic connection to the passage than the easy distractors.

Table 18 presents a longer passage from the CLOTH-H subset, which is based on high school English examinations. Similar to the previous case, a clear distinction in difficulty levels can be observed among the distractors. The augmented hard distractors contain words related to the passage's theme of holiday experiences, whereas the augmented easy distractors, though grammatically correct, are semantically unrelated to the main topic. A similar pattern emerges in the generated distractors, where the hard distractors include words associated with holidays and time, while the easy distractors mostly contain less relevant terms.

Passage	Yesterday I was tidying up my room. I found an old box of my father's. He gave it to me two years ago. It was fascinating to discover some of my father's childhood photos. He once told me that he wrote to people all over the world, and they sent him letters, too. As a result, he had a book of interesting stamps. People also gave him things from different countries, such as a silk from Japan, a little doll from England, and a small model ship from Australia. My father even kept he tickets from his first football match! It made me think about looking after my collection of _____ pictures books and magazines.
Answer	different

	Original	**Augmented**		**Generated**	
		Easy	**Hard**	**Easy**	**Hard**
Distractors	new	disappointing	old	difficult	old
	old	difficult	interesting	easy	beautiful
	interesting	disgusting	travel	funny	interesting
		bad	colorful	expensive	new
		cheap	school	small	other
		comic	many	common	various
		boring	stamps	same	good
		funny	useful	bad	similar

Table 17: An example of augmented and generated distractors.

Passage	When I was a boy, every holiday that I had seemed wonderful. My parents took me by train or by car to a hotel by the sea. All day, I seem to remember, I played on the sands with strange excited children. We made houses and gardens, and watched the tide destroy them. When the tide went out, we climbed over the rocks and looked down at the fish in the rock-pools. In those days the sun seemed to shine always brightly and the water was always warm. Sometimes we left the beach and walked in the country, exploring ruined houses and dark woods and climbing trees. There were sweets in one's pockets or good places where one could buy ice-creams. Each day seemed a life-time. Although I am now thirty-five years old, my idea of a good _____ is much the same as it was. I still like the sun and warm sand and the sound of waves beating the rocks. I no longer wish to build any sand house or sand garden, and I dislike sweets. However, I love the sea and often feel sand running through my fingers. Sometimes I wonder what my ideal holiday will be like when I am old. All I want to do then, perhaps, will be to lie in bed, reading books about children who make houses and gardens with sands, who watch the incoming tide, who make themselves sick on too many ices.
Answer	holiday

	Original	**Augmented**		**Generated**	
		Easy	**Hard**	**Easy**	**Hard**
Distractors	house	house	day	party	day
	garden	impression	summer	festival	time
	tide	job	time	week	trip
		month	fun	job	weekend
		game	experience	work	life
		event	weekend	dream	week
		evening	pleasure	meal	place
		night	life	dinner	vacation

Table 18: An example of augmented and generated distractors.