DualReward: A Dynamic Reinforcement Learning Framework for Cloze Tests Distractor Generation

Tianyou Huang[1,a], Xinglu Chen[1,a], Jingshen Zhang[1], Xinying Qiu[1,*], Ruiying Niu[2]
[1]Department of Computer Science, School of Information Science and Technology,
[2]Faculty of English Language and Culture
Guangdong University of Foreign Studies, Guangzhou, China
tenny314159@gmail.com, luc88749@gmail.com, audbut0702@163.com,
xy.qiu@foxmail.com, 199610539@gdufs.edu.cn

Abstract

This paper introduces DualReward, a novel reinforcement learning framework for automatic distractor generation in cloze tests. Unlike conventional approaches that rely primarily on supervised learning or static generative models, our method employs a dual reward structure with adaptive scaling that differentiates between human-created gold standard distractors and model-generated candidates. The framework dynamically adjusts reward signal intensity based on model performance and confidence. We evaluate our approach on both passage-level (CLOTH-F) and sentence-level (MCQ) cloze test datasets, demonstrating consistent improvements over state-of-the-art baselines. Experimental results show that our adaptive reward scaling mechanism provides modest but consistent benefits on homogeneous datasets (CLOTH-F) and more substantial improvements (3.48-3.86% in P@1) on diverse, cross-domain data (MCQ), suggesting its particular effectiveness for handling varied question types and domains. Our work offers a flexible framework that effectively balances learning from reliable human examples while exploring novel, high-quality distractors for automated test generation.

Keywords: Distractor Generation , Cloze Test , Reinforcement Learning

1 Introduction

Educational assessments play a critical role in evaluating student knowledge and comprehension, with cloze tests (fill-in-the-blank questions) being widely used due to their effectiveness in measuring language proficiency and content understanding. Automatic distractor generation (ADG) has emerged as a promising research direction to address the challenge of creating effective educational assessments, with diverse approaches evolving rapidly in recent years (Alhazmi et al., 2024). Earlier methods in this field employed knowledge-driven approaches that leverage semantic hierarchies and taxonomies to identify plausible distractors (Ren and Zhu, 2021). The current state-of-the-art methods have shifted toward candidate generating-and-ranking (CGR) frameworks using pre-trained language models (Chiang et al., 2022) and transformer-based architectures that can capture contextual information (Wang et al., 2023). Recent innovations by (Taslimipoor et al., 2024) have further demonstrated the effectiveness of combining generative and discriminative capabilities of transformer models to produce high-quality distractors across various question types.

When generating distractors automatically, models often produce mixed-quality candidates. As the example shown in Table 1, for the question "If an object is attracted to a magnet, the object is most likely made of ___" (answer: "metal"), a model might generate both effective distractors like "wood" and "plastic" alongside problematic ones like "steel" and "iron" (which are types of metals). Rather than discarding imperfect candidates, we propose learning from this full spectrum by differentially rewarding distractors based on their quality. This approach helps the model distinguish between pedagogically effective options and those that are semantically related but contextually inappropriate for assessment purposes.

[a]Equal Contribution; [*]Corresponding author

©2025 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Cloze Test Sentence	If an object is attracted to a magnet, the object is most likely made of ____.
Correct Answer	metal
Gold Standard Distractors	wood, plastic, cardboard
Generated Distractors	wood, plastic, glass, stone, steel, aluminum, clay, iron, rubber

Table 1: Distractors Generation Example

We introduce DualReward, a novel reinforcement learning framework for automatic distractor generation that addresses key limitations of existing approaches through two primary innovations. First, we propose a dual reward structure that differentiates between gold standard human-created distractors and model-generated candidates, enabling the model to learn from both reliable examples and its own generations while maintaining appropriate distinctions between them. Second, we develop an adaptive reward scaling mechanism that dynamically adjusts the intensity of learning signals based on model performance, providing stronger guidance when the model struggles with difficult examples and more subtle refinement as performance improves.

Our approach draws conceptual inspiration from recent advances in reinforcement learning, particularly adaptive reward mechanisms (Haarnoja et al., 2018) and curriculum learning (Florensa et al., 2018), while applying these principles to the unique challenges of educational content generation. Unlike previous approaches that treat all distractors uniformly, our method acknowledges the inherent difference in reliability between human-created and model-generated options, applying differential rewards scaled by model confidence.

We evaluate our DualReward framework on two established benchmarks: CLOTH-F (Wang et al., 2023), a passage-level cloze dataset, and MCQ (Ren and Zhu, 2021), a diverse cross-domain sentence-level dataset. Our results demonstrate that the proposed approach consistently outperforms state-of-the-art baselines across multiple metrics, with particularly notable improvements in ranking quality as measured by Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG).

Our contributions can be summarized as follows:

- We introduce a dual reward structure for distractor generation that effectively differentiates between gold standard and model-generated distractors, enabling more nuanced learning from both sources.
- We develop an adaptive reward scaling mechanism that dynamically adjusts learning signals based on model performance, facilitating more effective training without extensive hyperparameter tuning.
- We demonstrate through comprehensive experiments that our DualReward framework achieves state-of-the-art performance on both passage-level and sentence-level cloze tests, with significant improvements in distractor ranking quality.

Our work bridges the gap between recent advances in reinforcement learning and automated educational assessment, offering a flexible framework that can be applied to diverse cloze test formats and domains. We provide our codes at https://github.com/tenny314159/dualreward.

2 Related Research

Our work on adaptive reward-scaled reinforcement learning for distractor generation builds upon two primary research areas: (1) automatic distractor generation for educational assessments and (2) adaptive reward mechanisms in reinforcement learning.

2.1 Automatic Distractor Generation

Automatic distractor generation (ADG) has become increasingly important in educational technology to reduce the manual effort in creating high-quality assessment items (Alhazmi et al., 2024). Several

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122–1135, Jinan, China, August 11–14, 2025.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

approaches have been proposed for multiple-choice questions, ranging from context-sensitive inference rules (Zesch and Melamud, 2014) to difficulty-aware generation (Yeung et al., 2019).

Knowledge-driven methods (Ren and Zhu, 2021) leverage external knowledge bases to identify semantically related but incorrect options. Transformer-based approaches employ both generative and discriminative capabilities to produce plausible distractors (Taslimipoor et al., 2024). Chung et al. (2020) proposed a BERT-based distractor generation scheme with multi-tasking and negative answer training strategies.

For cloze tests specifically, several specialized methods have been developed. Chiang et al.(2022) introduced CDGP, an automatic cloze distractor generation approach based on pre-trained language models. Panda et al. (2022) explored round-trip neural machine translation for generating distractors in fill-in-the-blank exercises. Jiang & Lee (2017) focused on Chinese fill-in-the-blank items, while Zhang et al. (2023) introduced ClozEx for generating explanations for English cloze tests.

Recent advancements include DiVERT (Fernandez et al., 2024), which generates distractors with variational errors for mathematics questions, and approaches leveraging large language models with counterfactual contrastive decoding (Qu et al., 2024) and knowledge graph integration (Yu et al., 2024). Kumar et al. (2023) proposed a novel approach incorporating semantic similarity measures and contextual understanding to create effective distractors.

2.2 Adaptive Reward Mechanisms in Reinforcement Learning

Our reinforcement learning component builds on adaptive reward mechanisms developed in recent years. Haarnoja et al.(2018) introduced Soft Actor-Critic with automatically adjusted temperature parameters, conceptually related to our adaptive reward scaling. This approach optimizes entropy-regularized policies, balancing exploration and exploitation, similar to how our method balances confidence in generated distractors.

Policy gradient methods with dynamic reward scaling Schulman et al. (2017) provide a foundation for our approach by demonstrating how adaptive scaling factors can stabilize training. Hadfield-Menell et al.(2017) explored inverse reward design, highlighting the importance of dynamic reward structures in complex learning environments. Their work on reward uncertainty parallels our confidence-weighted approach.

In practice-oriented research, Wang et al. (2020) proposed methods for handling noisy rewards that share conceptual similarities with our approach to handling uncertain distractor quality. Florensa et al. (2018) introduced reverse curriculum generation, where task difficulty progressively increases as the agent improves, conceptually similar to our adaptive scaling mechanism that adjusts based on model performance.

Kwon et al.(2024) presented adaptive reward design for reinforcement learning in complex robotic tasks with an emphasis on dynamic difficulty adjustment. In the context of curriculum learning, Narvekar et al. (2020) reviewed approaches where task difficulty adapts based on learner progress, paralleling our method of adjusting reward scaling based on model performance.

Meta-gradient reinforcement learning (Xu et al., 2018) provides theoretical foundations for hyperparameter adaptation that inspired our sigmoid-based reward scaling mechanism. Similarly, Zheng (2018) developed methods for learning intrinsic reward functions that can be dynamically adapted during training, which informed our approach to automatically adjusting reward scales based on training progress.

Our work bridges these research areas by introducing an adaptive reward-scaled reinforcement learning framework specifically designed for distractor generation. Unlike previous approaches that rely primarily on supervised learning or static reward structures, our method dynamically adjusts reward signals based on model performance and confidence, with a dual reward structure that differentiates between gold standard and model-generated distractors.

3 Methodology

Our approach addresses automatic distractor generation for cloze tests using a adaptive reward-scaled reinforcement learning framework. Figure 1 provides an overview of our approach using a concrete

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122–1135, Jinan, China, August 11–14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

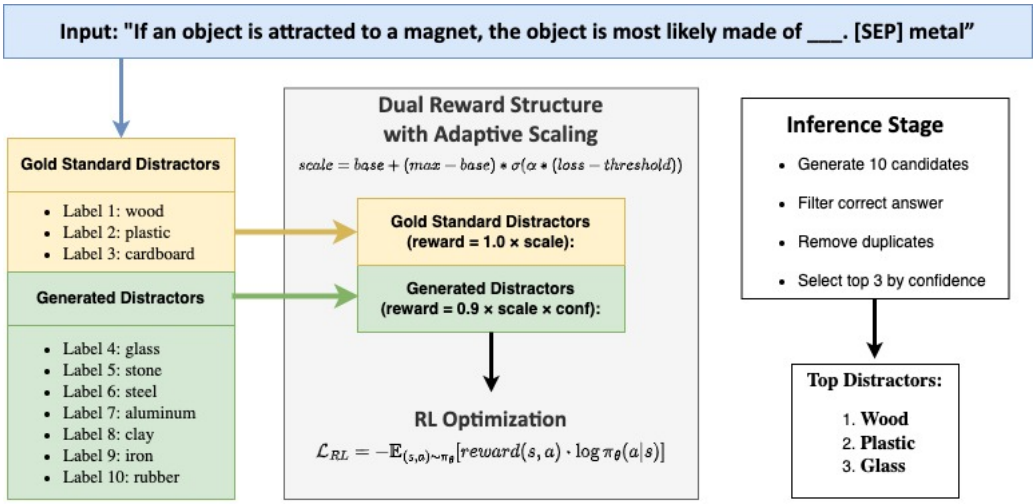

Figure 1: DualReward Framework

example: for the cloze test "If an object is attracted to a magnet, the object is most likely made of ____" with the correct answer "metal", our framework processes this input through four key components: (1) a distractor candidate generation model that produces gold standard distractors (wood, plastic, cardboard) and model-generated candidates (glass, stone, steel, aluminum, clay, iron, rubber), (2) an adaptive reward scaling mechanism, (3) a dual reward structure that differentiates between these two types of distractors, and (4) an inference procedure that selects optimal distractors based on confidence scores.

3.1 Training Data Generation

Our distractor generation process consists of the following steps:

- Initial Candidate Generation: For each cloze test instance, we generate an initial pool of candidate distractors using our base language model.
- Candidate Filtering: We filter out candidates that are identical to any gold standard distractors from the training set, ensuring the model learns to generate novel distractors rather than memorizing examples.
- Supplementary Generation: If fewer than 7 unique candidates remain after filtering, we perform additional generation passes until we obtain 7 distinct candidate distractors.
- Training Data Preparation: For each training instance, we compile a set of 10 potential answers (labels): 3 gold standard distractors from human-created references and 7 model-generated candidate distractors.

We conducted pilot study to empirically test different pool sizes of candidate distractors, including 10 generated candidates + 3 gold standar (total 13). Results remained similar, so we chose 7 generated + 3 gold standard for a balanced total of 10 candidates.

3.2 Dual Reward Structure with Adaptive Scaling

The core of our approach lies in a dual reward structure with adaptive scaling that differentiates between gold standard distractors and model-generated candidates. Our framework draws conceptual inspiration from asymmetric learning paradigms in other domains. For instance, Zhang and Lapata (2017) used asymmetric rewards in sentence simplification. However, we design differential rewards between gold-standard and model-generated distractors. Human-created distractors are treated as fully reliable examples. We use the Candidate Set Generation (CSG) module from the CDGP model by Chiang et al. (2022) to generate candidate distractor sets. In CDGP, each candidate distractor receives a

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122–1135, Jinan, China, August 11–14, 2025.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$confidence_score$ from BERT's masked language modeling head, computed as the softmax probability of the distractor token at the masked position. Model-generated distractors receive a discounted reward based on model confidence. The discount factor reflects inherent uncertainty in model-generated versus human-validated content. The 0.9 value provides meaningful but not excessive penalty, encouraging exploration while preventing overconfidence in generated distractors.

Gold Standard Distractors:

$$reward_{gold} = 1 \times reward_scale$$

Generated Candidate Distractors:

$$reward_{gen} = 0.9 \times reward_scale \times confidence_score$$

where $confidence_score$ represents the model's confidence in the candidate distractor's suitability.

Adaptive Reward Scaling

Our adaptive reward scaling mechanism builds upon principles established in reinforcement learning literature. The sigmoid-based transition function resembles the temperature adaptation in Soft Actor-Critic (Haarnoja et al., 2018), where the temperature parameter dynamically adjusts to balance exploration and exploitation. The specific form of our scaling equation:

$$reward_scale = base_scale + (max_scale - base_scale) \cdot \sigma(\alpha \cdot (avg_loss - threshold))$$

where:

- $base_scale$ represents the minimum reward scaling factor
- max_scale represents the maximum reward scaling factor
- α controls the steepness of the transition between scaling factors
- $threshold$ defines the loss value at which the scaling begins to transition
- σ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$
- avg_loss is the average loss during the current training phase

The dual rewards establish a hierarchy between gold-standard and generated distractors, while the adaptive scaling dynamically adjusts the overall intensity of learning signals based on model performance. This integrated approach allows the model to simultaneously learn from reliable examples while exploring the generation of novel distractors, with the learning signal strength automatically calibrated to the current stage of training.

We optimize the model using a reinforcement learning objective that incorporates these scaled rewards:

$$\mathcal{L}_{RL} = -\mathbb{E}_{(s,a)\sim\pi_\theta}[reward(s,a) \cdot \log \pi_\theta(a|s)]$$

where π_θ represents our policy (the distractor generation model), s is the input sentence, and a is the generated distractor.

3.3 Inference Procedure

During test time, we employ the following procedure to select the final distractors:

- Generate 10 candidate distractors for the given cloze test
- Filter out candidates that are identical to the correct answer
- Remove duplicate candidates
- Select the top 3 candidates based on the model's confidence scores

This approach ensures that the final distractors are diverse, distinct from the correct answer, and optimize the model's learned criteria for effective distractors.

CCL 2025

4 Experiment Design

4.1 Dataset

We evaluate our DualReward framework on two established benchmarks for cloze test distractor generation. **CLOTH-F (Filtered)** is a refined subset of the CLOTH dataset, introduced by (Wang et al., 2023). The original CLOTH dataset contains questions with numbered blanks (e.g., ”1”). To avoid training data inconsistency and create a more standardized evaluation setting, CLOTH-F removes the questions with numbered blanks, resulting in 5,041/720/739 instances for train/dev/test splits. Each instance consists of a passage-level cloze test with a blank indicated by an underscore, a correct answer, and human-generated distractors.

MCQ is a cross-domain, sentence-level cloze dataset spanning science, vocabulary, common sense, and trivia domains. It contains 2,321/258 questions for train/test splits (with dev data created by a 9:1 split from training), where each instance has a sentence-level context with blanks marked as blank. While CLOTH-F provides rich contextual information through full passages, MCQ offers domain diversity with more focused sentence-level questions, allowing us to evaluate our model's performance across different cloze test formats and difficulty levels. Table 2 provides the dataset statistics.

Dataset	CLOTH-F				MCQ			
	Train	Dev	Test	All	Train	Dev	Test	All
# of Passages	5,041	720	739	6500	-	-	-	-
# of Questions	69,009	9,696	10,233	88,938	2088	233	258	2580

Table 2: Statistics for the training, development, and test sets of CLOTH-F and MCQ.

4.2 Baseline Models

We compare with the following baselines on experiments with CLOTH-F dataset:

T5 multi-task (+ CTA): (Wang et al., 2023) enhance T5 distractor generation through multi-task training with cloze test answering (CTA). The model jointly learns to generate distractors and answer cloze questions, where CTA requires selecting the correct answer from shuffled options. This dual-task approach improves the model's understanding of answer-distractor relationships.

T5 multi-task (+ DF, CTA): (Wang et al., 2023) extend their multi-task framework by adding distractor finding (DF) to CTA. The DF task trains the model to identify distractors placed in cloze gaps, creating a three-task training setup with balanced data sampling. This comprehensive approach achieves their best results on CLOTH-F with 28.47 P@1.

We compare with the following baselines on experiments with MCQ dataset:

T5 multi-task (+DF): (Wang et al., 2023) enhances T5 distractor generation through multi-task training with the distractor finding (DF) task. The DF task trains the model to identify distractor spans within cloze passages by placing a distractor at the blank position and having the model generate that distractor based on the modified input. This additional task helps the model better understand distractor characteristics.

CSG-DS: (Ren and Zhu, 2021) employs a candidate generating-and-ranking framework consisting of two stages: a Candidate Set Generator (CSG) that extracts semantically similar candidates from knowledge bases using context-dependent conceptualization, and a Distractor Selector (DS) that re-ranks candidates using a learning-to-rank model with elaborately designed features measuring both plausibility and reliability.

CDGP (Chiang et al., 2022) leverages pre-trained language models (specifically BERT) for distractor generation within the candidate generating-and-ranking paradigm. The model fine-tunes BERT to predict distractors from the masked blank position, using the language model's inherent fill-in-the-blank capability from masked language modeling pretraining.

CCL 2025

RAP-Pt5: (Yu et al., 2024) introduces Retrieval Augmented Pretraining for the DG task. This approach uses MCQ answers to retrieve relevant sentences/passages from large corpora like Wikipedia to create pseudo questions for task-specific pretraining. In this specific implementation, the T5 model is first pretrained on the Sciq-all dataset using RAP before being fine-tuned for MCQ distractor generation, providing cross-domain knowledge transfer from the larger Sciq dataset to improve MCQ performance.

4.3 Implementation Details

For the initial candidate generation phase, we apply BERT within the Candidate Set Generation (CSG) framework established by (Chiang et al., 2022). We employ T5-base as our foundational language model for distractor selection experiments. When conducting experiments on the MCQ dataset, we first pretrain the T5-base model on the Sciq-all dataset before implementing our DualReward framework, enabling cross-domain knowledge transfer from the larger scientific corpus following (Yu et al., 2024).

Our hyperparameter configuration includes 20 training epochs with a batch size of 8, utilizing a learning rate of 1e-4 with the AdamW optimizer. Our adaptive scaling provides stronger reinforcement signals when the model struggles (high loss) and more subtle guidance as performance improves. We set $base_scale = 0.1$ and $max_scale = 0.2$ to provide a controlled range of reward magnitudes, following similar bounded scaling approaches in policy gradient methods (Schulman et al., 2017). The steepness parameter $\alpha = 5$ was selected to create a responsive yet stable transition, similar to sensitivity parameters in curriculum learning (Florensa et al., 2018). The threshold = 1.0 was determined based on preliminary experiments, representing an inflection point in learning dynamics where model confidence begins to stabilize.

For comprehensive evaluation, we adopt the standard metrics established by state-of-the-art models in distractor generation. These include Precision@1 (P@1), which measures the accuracy of the top-ranked distractor; Recall@1 (R@1), which evaluates the proportion of ground-truth distractors retrieved at rank one; F1@3, representing the harmonic mean of precision and recall for the top three predictions; Mean Reciprocal Rank@3 (MRR@3), which captures the ranking quality by computing the reciprocal of the rank at which the first correct distractor appears; and Normalized Discounted Cumulative Gain@3 (NDCG@3), which provides a graded relevance assessment with position-based discounting. Together, these metrics offer a comprehensive view of both generation quality and ranking effectiveness.

5 Results and Analysis

5.1 Main Results

The experimental results in Table 3 demonstrate the effectiveness of our proposed DualReward framework across both datasets. On the CLOTH-F dataset, our approach achieves state-of-the-art performance with a P@1 of 28.87%, slightly outperforming the strong T5 multi-task baseline methods. While the absolute improvement in precision metrics is modest (0.4% over the CTA baseline), our method shows more substantial gains in ranking quality metrics, with a 2.06% improvement in MRR@3 and a remarkable 6.56% gain in NDCG@3 compared to the best baseline.

These improvements in ranking metrics are particularly significant as they indicate that our method not only identifies the most plausible distractor correctly but also produces a more effective overall ranking of distractor candidates. The F1@3 score of 21.87% represents a 2.05% absolute improvement over the best baseline, demonstrating better precision-recall balance across the top three predictions.

On the MCQ dataset, the advantages of our approach are even more pronounced. The DualReward framework with pretraining achieves 37.84% P@1, surpassing the strong RAP-PT5 baseline by 6.18%. This substantial improvement is consistent across all metrics, with particularly notable gains in MRR@3 (8.5% improvement) and NDCG@3 (14.23% improvement). These results highlight that our adaptive reward scaling mechanism is especially effective in the cross-domain setting with sentence-level cloze tests, where the model must handle diverse topics and question formats.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122-1135, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Datasets	Model	P@1	R@1	F1@3	MRR@3	NDCG@3
CLOTH-F	T5 multi-task (+ CTA)	28.75	9.58	19.20	34.06	35.64
	T5 multi-task (+ DF, CTA)	28.47	9.49	19.82	34.46	36.26
	Dual-Reward	**28.87**	**9.63**	**21.87**	**36.52**	**42.82**
MCQ	T5 multi-task (+ DF)	22.00	7.33	13.64	27.15	28.50
	CSG+DS	10.08	-	9.42	17.33	-
	CDGP	13.13	-	12.23	-	-
	RAP-PT5 (w/ pretrain)	31.66	10.55	18.91	37.70	39.43
	Dual-Reward (w/pretrain)	**37.84**	**12.61**	**23.42**	**46.20**	**53.66**

Table 3: Main Results

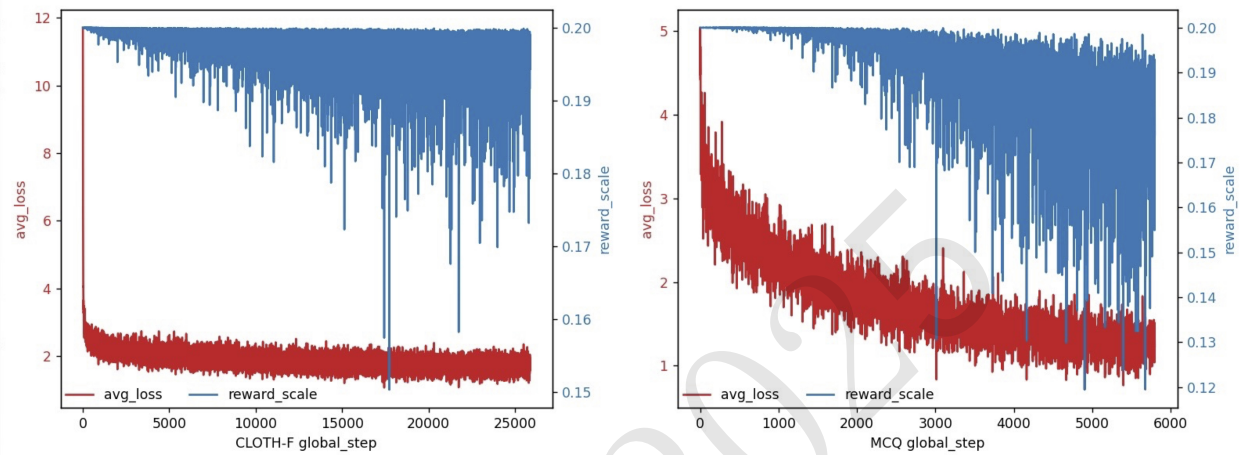

Figure 2: Change of Reward Scale

5.2 Reward Scale

Figure 2 illustrates the dynamic nature of our adaptive reward scaling mechanism during training on both CLOTH-F and MCQ datasets. As shown in the figure, the reward scale (right y-axis, blue line) changes in response to the average loss (left y-axis, red line) throughout the training process. For the CLOTH-F dataset, we observe that the reward scale starts at a higher value and gradually decreases as the model's performance improves and the loss decreases. This pattern demonstrates how our mechanism provides stronger learning signals during the initial training phases when the model is less confident, and then transitions to more subtle guidance as performance stabilizes. The MCQ dataset shows a similar but more volatile pattern, with the reward scale fluctuating more frequently in response to the diverse, cross-domain nature of the data. These visualizations support our quantitative findings, confirming that adaptive scaling effectively responds to the model's changing learning dynamics, particularly when handling varied question types and domains.

5.3 Ablation Study

5.3.1 Dual Reward vs. Uniform Reward

To further validate the importance of our adaptive reward scaling mechanism, we conducted an ablation study where we replaced the dynamic reward scale with constant values. This differs from our dual reward approach where model-generated distractors receive a discounted reward ($0.9 \times reward_scale \times confidence_score$).

$$reward_{gold} = 1 \times reward_scale$$

$$reward_{gen} = 1 \times reward_scale \times confidence_score$$

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

As shown in Table 4, the dual reward structure consistently outperforms the uniform reward approach. On CLOTH-F, our method achieves a 0.53% improvement in P@1 (28.87% vs. 28.34%) and more substantial gains in ranking metrics (0.66% in MRR@3, 0.71% in NDCG@3). For the MCQ dataset, the improvements are more pronounced, with a 4.64% gain in P@1 (37.84% vs. 33.20%) and similar improvements in ranking metrics. These results confirm that establishing a clear hierarchy between human-created examples and model-generated candidates enables more effective learning, particularly in diverse, cross-domain settings like the MCQ dataset.

5.3.2 Adaptive Reward vs. Constant Reward

To evaluate our adaptive reward scaling mechanism, we compared it against three fixed reward scale values (0.1, 0.15, and 0.2) while maintaining the dual reward structure. This eliminates the adaptive component that dynamically adjusts based on model performance during training:

$$reward_{gold} = 1 \times constant_reward$$

$$reward_{gen} = 0.9 \times constant_reward \times confidence_score$$

As shown in Table 4, the adaptive approach consistently outperforms fixed scaling across all metrics on both datasets. On CLOTH-F, the improvements are modest but consistent, with gains of 0.21-0.46% in P@1 over the constant reward baselines. The MCQ dataset shows more substantial benefits from adaptive scaling, with improvements of 3.48-3.86% in P@1 compared to the constant reward variants.

The more pronounced improvements on MCQ suggest that adaptive scaling is particularly beneficial for diverse, cross-domain datasets where learning dynamics may vary more substantially during training. When encountering questions from unfamiliar domains, the adaptive mechanism provides stronger learning signals, facilitating more effective transfer learning.

Analysis of the results across constant reward settings reveals that no single fixed value works optimally across different datasets or metrics, underscoring the value of our adaptive approach, which automatically calibrates reward intensity based on the model's current learning state without requiring extensive hyperparameter tuning. We also performed an ablation study on the MCQ dataset where we did not pretrain T5-base model on the Sciq-all dataset before implementing our DualReward framework. The results as shown in the last row of Table 4 are lower than the Dual-Reward with pretrain.

Datasets	Model	P@1	R@1	F1@3	MRR@3	NDCG@3
CLOTH-F	Dual-Reward	**28.87**	**9.63**	**21.87**	**36.52**	**42.82**
	w/ constant reward = 0.1	28.66	9.55	21.55	36.09	42.21
	w/ constant reward = 0.15	28.41	9.47	21.65	36.01	42.31
	w/ constant reward = 0.2	28.56	9.52	21.69	36.21	42.58
	Uniform-Reward	28.34	9.45	21.38	35.86	42.11
MCQ	Dual-Reward (w/pretrain)	**37.84**	**12.61**	**23.42**	**46.20**	**53.66**
	w/ constant reward = 0.1	33.98	11.33	23.17	42.92	50.49
	w/ constant reward = 0.15	34.36	11.45	23.29	43.82	52.19
	w/ constant reward = 0.2	33.98	11.33	21.88	42.34	49.59
	Uniform-Reward	33.20	11.07	23.42	42.54	50.75
	Dual-Reward (w/o pretrain)	22.39	7.46	13.64	27.86	32.53

Table 4: Ablation Study of Adaptive Reward Scale and Uniform Reward

5.4 Qualitative Analysis

To gain deeper insights into the effectiveness of our DualReward framework, we present sample outputs in Tables 5 and 6 for MCQ and CLOTH-F datasets, respectively. These examples demonstrate how our approach improves distractor quality compared to the Uniform-Reward baseline.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In the MCQ examples (Table 5), our DualReward framework generates more contextually appropriate distractors. For instance, in Sentence 1 regarding transverse waves, our model produces ”peaks, ridges, valleys” versus the baseline’s ”plateaus, peaks, valleys,” where ”plateaus” is less appropriate for wave physics terminology. In Sentence 4 about photosynthesis, our model generates ”carbon, helium, hydrogen” while the baseline includes ”water,” which is problematic since water is actually a photosynthesis reactant.

For the CLOTH-F passage (Table 6), our framework demonstrates better contextual awareness. In blank 1 where the answer is ”eating,” our model generates ”working, studying, thinking” - all plausible breakfast meeting activities, while the baseline produces ”sleeping, playing, studying” where ”sleeping” is contextually inappropriate. Similarly, for blank 11 (answer: ”decision”), our model generates ”plan, promise, suggestion” - all commitment-related terms, versus the baseline’s inclusion of ”explanation” which fits less well in the narrative context.

While our DualReward framework generally produces better distractors, some limitations remain. In MCQ Sentence 7, our model generates ”leaves, stems, seeds” for gymnosperms that ”do not have flowers,” where ”seeds” is problematic since gymnosperms actually do have seeds. Similarly, in CLOTH-F blank 10 (answer: ”hit”), our model produces ”beat, pleased, attracted” where ”beat” could be considered too close in meaning to the correct answer.

These examples demonstrate that while our DualReward framework generally improves distractor quality through better contextual appropriateness, challenges remain in avoiding semantic overlaps and factual inconsistencies.

6 Conclusion

In this paper, we introduced DualReward, a novel reinforcement learning framework for automatic cloze test distractor generation that employs both adaptive reward scaling and a dual reward structure. Our approach dynamically adjusts learning signals based on model performance while differentiating between gold standard and model-generated distractors, addressing key limitations in existing supervised learning approaches. Experimental results on both CLOTH-F and MCQ benchmarks demonstrate consistent improvements over state-of-the-art approaches, with particularly notable gains on the diverse MCQ dataset (6.18% improvement in P@1, 8.5% in MRR@3, and 14.23% in NDCG@3). Comprehensive ablation studies confirm the benefits of our adaptive reward scaling mechanism, especially in cross-domain settings where learning dynamics vary substantially during training.

While our evaluation primarily focuses on matching human-created distractors rather than directly measuring pedagogical effectiveness, the DualReward framework provides a promising foundation for balancing learning from human-created examples while generating novel distractors. Future research directions include extending the framework to multilingual cloze tests, incorporating pedagogically-aligned evaluation metrics that measure actual learning outcomes, and exploring applications to other educational assessment formats such as reading comprehension and mathematical problem generation. The adaptive reward mechanism also holds potential for broader applications in other text generation tasks requiring quality differentiation between training examples.

7 Limitations

Despite promising results, our approach has several limitations. First, our evaluation focuses on matching human-created distractors rather than measuring actual educational effectiveness, leaving pedagogical impact incompletely assessed. Second, experiments are limited to English cloze tests due to the absence of established cross-lingual benchmarks in this domain. Future work should incorporate pedagogically-aligned evaluation metrics that measure learning outcomes and extend to multilingual settings as appropriate datasets become available.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122-1135, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

References

Alhazmi, E., Q. Z. Sheng, W. E. Zhang, M. Zaib, and A. Alhazmi. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *arXiv preprint arXiv:2402.01512*.

Chiang, S.-H., S.-C. Wang, and Y.-C. 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.

Chung, H. L., Y. H. Chan, and Y. C. Fan. 2020, November. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4390–4400.

Fernandez, N., A. Scarlatos, W. Feng, S. Woodhead, and A. Lan. 2024. DiVERT: Distractor generation with variational errors represented as text for math multiple-choice questions. *arXiv preprint arXiv:2406.19356*.

Florensa, C., D. Held, X. Geng, and P. Abbeel. 2018, July. Automatic goal generation for reinforcement learning agents. In *International Conference on Machine Learning*, pages 1515–1528. PMLR.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Hadfield-Menell, D., S. Milli, P. Abbeel, S. J. Russell, and A. Dragan. 2017, December. Inverse reward design. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Jiang, S. and J. S. Lee. 2017, September. Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Kumar, A. P., A. Nayak, M. S. K., S. Goyal, and Chaitanya. 2023. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225:120022.

Kwon, M., I. ElSayed-Aly, and L. Feng. 2024. Adaptive reward design for reinforcement learning in complex robotic tasks. *arXiv preprint arXiv:2412.10917*.

Narvekar, S., B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50.

Panda, S., F. P. Gomez, M. Flor, and A. Rozovskaya. 2022, May. Automatic generation of distractors for fill-in-the-blank exercises with round-trip neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401.

Qu, F., H. Sun, and Y. Wu. 2024. Unsupervised distractor generation via large language model distilling and counterfactual contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 827–838.

Ren, S. and K. 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.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Taslimipoor, S., L. Benedetto, M. Felice, and P. Buttery. 2024, May. Distractor generation using generative and discriminative capabilities of transformer-based models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5052–5063.

Wang, J., Y. Liu, and B. Li. 2020, April. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, number 04, pages 6202–6209.

Wang, H. J., K. Y. Hsieh, H. C. Yu, J. C. Tsou, Y. A. Shih, C. H. Huang, and Y. C. Fan. 2023, July. 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.

Xu, Z., H. P. van Hasselt, and D. Silver. 2018. Meta-gradient reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 2396–2407.

Yeung, C. Y., J. S. Lee, and B. K. 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.

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

Zesch, T. and O. Melamud. 2014, June. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Zhang, X. and M. Lapata. 2017, September. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 584–594. Association for Computational Linguistics.

Zheng, Z., J. Oh, and S. Singh. 2018, December. On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*.

Zhang, Z., M. Mita, and M. Komachi. 2023. ClozEx: A task toward generation of English cloze explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5228–5242.

CCL 2025

8 Appendix

Sentence	1. The high points of a transverse wave are called ___ .
Answer / Distractor	crests / peaks, troughs, ridges
Dual-Reward	peaks, ridges, valleys
Uniform-Reward	plateaus, peaks, valleys
Sentence	2. In sexual reproduction , ___ is the name of the gamete cell the male must contribute.
Answer / Distractor	sperm / plasma, spore, ova
Dual-Reward	egg, dna, fetus
Uniform-Reward	egg, dna, rna
Sentence	3. The main organs of the respiratory system are ___ .
Answer / Distractor	lungs / ovaries, intestines, kidneys
Dual-Reward	ovaries, kidneys, liver
Uniform-Reward	ovaries, kidney, liver
Sentence	4. The products of photosynthesis are glucose and ___ else .
Answer / Distractor	oxygen / carbon, hydrogen, nitrogen
Dual-Reward	carbon, helium, hydrogen
Uniform-Reward	nitrogen, hydrogen, water
Sentence	5. ___ is responsible for erosion by flowing water and glaciers .
Answer / Distractor	gravity / kinetic, electromagnetic, weight
Dual-Reward	weight, magnetism, electricity
Uniform-Reward	motion, friction, magnetism
Sentence	6. ___ of rocks form from cooled magma or lava .
Answer / Distractor	igneous / granite, sedimentary, metamorphic
Dual-Reward	metamorphic, sedimentary, meta
Uniform-Reward	metamorphic, granite, lava
Sentence	7. Gymnosperms have seeds but do not have ___ .
Answer / Distractor	flowers / leaves, stems, roots
Dual-Reward	leaves, stems, seeds
Uniform-Reward	leaves, stems, stem
Sentence	8. Some 96 % of the dry mass consists of organic compounds produced by ___ .
Answer / Distractor	photosynthesis / erosion, electrolysis, reproduction
Dual-Reward	spermatogenesis, glycolysis, erosion
Uniform-Reward	glycolysis, spermatogenesis, osmosis
Sentence	9. It is called ___ when something is unable to move from place to place .
Answer / Distractor	sessile / Undrate, obovate, ovate
Dual-Reward	elongated, spile, ovate
Uniform-Reward	spongile, elliptical, aqueous
Sentence	10. The greatest contribution of arthropods to human food supply is ___ .
Answer / Distractor	pollination / vegetation, reproduction, hibernation
Dual-Reward	migration, precipitation, vegetation
Uniform-Reward	migration, sedimentation, eva

Table 5: MCQ Generated Distractors Example

CCL 2025

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1122-1135, Jinan, China, August 11-14, 2025.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Passage
Years ago, a critical event occurred in my life that would change it forever. I met Kurt Kampmeir of Success Motivation Incorporation for breakfast. While we were 1 , Kurt asked me, " John, what is your 2 for personal growth?" Never at a loss for words, I tried to find things in my life that might 3 for growth. I told him about the many activities in which I was 4 . And I went into a 5 about how hard I worked and the gains I was making. I must have talked for ten minutes. Kurt 6 patiently, but then he 7 smiled and said, "You don't have a personal plan for growth, do you?" "No, I 8 ." "You know," Kurt said simply, "growth is not a(n) 9 process." And that's when it 10 me. I wasn't doing anything on purpose to make myself better. And at that moment, I made the 11 : I will develop and follow a personal growth plan for my 12 . That night, I talked to my wife about my 13 with Kurt and what I had learned. I 14 her the workbook and tapes Kurt was selling. We 15 that Kurt wasn't just trying to make a sale. He was offering a 16 for us to change our lives and achieve our dreams. Several important things happened that day. First, we decided to 17 the resources. But more importantly, we made a commitment to 18 together as a couple. From that day on, we learned together, traveled together, and sacrificed together. It was a 19 decision. While too many couples grow apart, we were growing together.

Answer / Distractor	Dual-Reward	Uniform-Reward
1. eating / working, preparing, thinking	1. working, studying, thinking	1. sleeping, playing, studying
2. plan / suggestion, demand, request	2. way, method, idea	2. dream, idea, way
3. qualify / appeal, look, call	3. appeal, provide, care	3. arrange, provide, care
4. involved / trapped, lost, bathed	4. occupied, attracted, invited	4. occupied, lost, included
5. speech / lecture, discussion, debate	5. lecture, discussion, report	5. discussion, joke, story
6. listened / calculated, drank, explained	6. cried, laughed, drank	6. cried, shouted, talked
7. finally / eagerly, gradually, gratefully	7. suddenly, luckily, immediately	7. suddenly, luckily, immediately
8. admitted / interrupted, apologized, complained	8. refused, agreed, argued	8. refused, agreed, argued
9. automatic / slow, independent, changing	9. natural, important, difficult	9. natural, ordinary, important
10. hit / confused, informed, pleased	10. beat, pleased, attracted	10. beat, attracted, comfort
11. decision/comment, announcement, arrangement	11. plan, promise, suggestion	11. promise, suggestion, explanation
12. life / progress, performance, investment	12. progress, study, family	12. family, career, study
13. conversation / contract, negotiation, argument	13. decision, competition, agreement	13. connection, competition, agreement
14. showed / lent, sold, offered	14. gave, lent, sent	14. sent, lent, brought
15. recognized / recalled, defined, declared	15. believed, hoped, imagined	15. hoped, imagined, guessed
16. way / tool, method, rule	16. method, chance, road	16. gift, road, place
17. buy / provide, give, deliver	17. save, use, provide	17. save, use, borrow
18. grow / survive, move, gather	18. work, study, gather	18. live, work, study
19. wise / difficult, random, firm	19. difficult, simple, common	19. difficult, simple, strange

Table 6: CLOTH - F Generated Distractors Example

CCL 2025

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China