

Automating True-False Multiple-Choice Question Generation and Evaluation with Retrieval-based Accuracy Differential

Chen-Jui Yu, Wen-Hung Lee, Lin-Tse Ke, Shih-Wei Guo, Yao-Chung Fan*

Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan
yfan@nchu.edu.tw

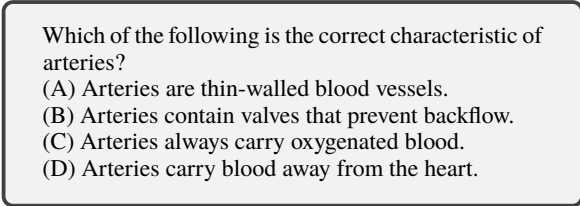
Abstract

Creating high-quality True-False (TF) multiple-choice questions (MCQs), with accurate distractors, is a challenging and time-consuming task in education. This paper introduces True-False Distractor Generation (TFDG), a pipeline that leverages pre-trained language models and sentence retrieval techniques to automate the generation of TF-type MCQ distractors. Furthermore, the evaluation of generated TF questions presents a challenge. Traditional metrics like BLEU and ROUGE are unsuitable for this task. To address this, we propose a new evaluation metric called Retrieval-based Accuracy Differential (RAD). RAD assesses the discriminative power of TF questions by comparing model accuracy with and without access to reference texts. It quantitatively evaluates how well questions differentiate between students with varying knowledge levels. This research benefits educators and assessment developers, facilitating the efficient automatic generation of high-quality TF-type MCQs and their reliable evaluation.

1 Introduction

Multiple-choice questions (MCQs) are an essential part of evaluative instruments for education. However, creating MCQs manually can be time-consuming and laborious. The core challenge part for MCQs' design is to craft *distractors* (wrong options). As a result, researchers have been working on automatic MCQ's distractor generation for different exam settings, such as reading comprehension (Chung et al., 2020; Gao et al., 2019), Cloze Quiz (Chiang et al., 2022; Yu et al., 2024), knowledge QA (Zhou et al., 2019).

Despite significant progress in the field, the generation of distractors for True-False (TF) MCQs has received limited attention. TF-type MCQs typically present four statement options, one correct and three incorrect, as shown in Figure 1, requiring



Which of the following is the correct characteristic of arteries?
(A) Arteries are thin-walled blood vessels.
(B) Arteries contain valves that prevent backflow.
(C) Arteries always carry oxygenated blood.
(D) Arteries carry blood away from the heart.

Figure 1: Example of True-False Type Multiple-choice Question

respondents to identify the correct option. These questions are commonly used in knowledge-based assessments, where participants must judge the accuracy of given statements.

However, there is a notable research gap in the automatic generation of TF-type distractors. While distractor generation has advanced in cloze tests (Liang et al., 2018; Yeung et al., 2019; Ren and Zhu, 2021; Chiang et al., 2022; Yu et al., 2024) and reading comprehension (Gao et al., 2019; Chung et al., 2020; Peng et al., 2022), the challenges of crafting true-false distractors remain underexplored. To address this, we introduce TFDG, a pipeline that integrates pre-trained language models and sentence retrieval techniques for True-False Distractor Generation.

Furthermore, a challenge in TFDG lies in the evaluation of its effectiveness. Traditional token-based metrics, like BLEU or ROUGE, do not quite encapsulate the essence of performance. These scores predominantly gauge n-gram overlap between the generated content and a reference. However, the essence of TF generation is not just about matching a reference but ensuring the crafted statements stand accurate and contextually relevant. While human evaluation, as utilized by (Zou et al., 2022), might seem a plausible route, it is not devoid of complications, such as potential subjectivity or varied review standards. As such, developing a robust evaluation metric for TF question generation presents another challenge.

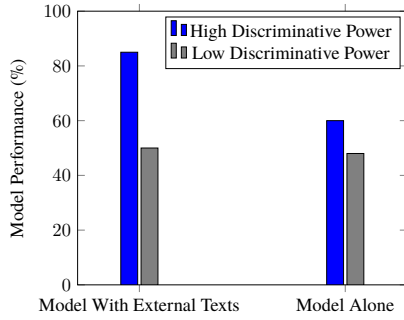


Figure 2: Evaluating MCQ Discriminative Power Using Model Performance Variations

To address this issue, we propose evaluating performance by considering the discriminative power of the questions, which refers to their ability to distinguish between respondents with varying levels of knowledge. A well-constructed multiple-choice question (MCQ) should exhibit high discriminative power, clearly differentiating between students who are familiar with the material and those who are not.

To facilitate this, we introduce the *Retrieval-based Accuracy Differential* (RAD), a metric that gauges the accuracy disparity of the MCQ answering model with and without access to reference texts. By comparing the model’s accuracy across these conditions using generated MCQs, we can discern how effectively the model can select the correct answer, thereby evaluating the discriminative power of the MCQs. This method highlights our ability to quantitatively measure the quality of MCQs, enhancing the robustness of MCQ evaluation.

To further illustrate this concept, consider a visual representation shown in Figure 2 comparing the behavior difference of MCQ answering models when faced with questions of varying discriminative power. In this depiction, the difference in model behavior, with and without access to external reference texts, can shed light on the discriminative power of the MCQs. MCQs with high discriminative power should show a significant divergence in the model’s behavior when external resources are either accessible or withheld, signifying that a well-crafted question can be resolved based on the prior knowledge provided. Conversely, for MCQs with low discriminative power, the model’s behavior is anticipated to remain consistent across both scenarios, suggesting that such questions might be too simplistic, ambiguous, or not thoroughly aligned with the tested content.

The contributions of this paper are as follows.

- We present TFDG, a pipeline that combines pre-trained language models and sentence retrieval techniques for True-False Distractor Generation.
- We present the RAD measure, the difference in accuracy of the MCQ answering model, measured with and without the provision of retrieval texts, to evaluate the performance of TFDG.

2 Related Work

In this section, we review the literature related to this work. Existing distractor generation (DG) methods can be broadly categorized into two main approaches: *cloze distractor generation* and *reading comprehension (RC) distractor generation*.

In the cloze DG task, the problem is approached as a word filling challenge. Typically, the first step involves extracting distractor candidates from the context or a knowledge base, followed by ranking the extracted distractors to produce the final result. Existing models in this field primarily rely on similarity heuristics (Guo et al., 2016; Ren and Q. Zhu, 2021) or supervised learning (Liang et al., 2018; Yeung et al., 2019; Ren and Zhu, 2021; Chiang et al., 2022).

On the other hand, the RC-type DG focuses on generating sentence-level distractors for reading comprehension level testing, such as summarizing an article or understanding author’s opinion (Gao et al., 2019; Zhou et al., 2019; Chung et al., 2020; Peng et al., 2022). For sentence-level distractor generation, neural models are commonly employed.

Delving into the available literature, the study by (Zou et al., 2022) emerges as closely aligned with our research aims. The authors introduce an unsupervised True/False Question Generation technique (TF-QG). Nevertheless, their methodology is tailored toward reading comprehension assessments intended for English learners. This deviates from our goal of crafting TF questions for knowledge-centric quizzes. As a result, there is a need to develop a new method for generating TF questions that is more aligned with our goal. Furthermore, in (Zou et al., 2022), performance evaluation was conducted through human evaluation. However, assessing the quality of a question through human evaluation can lead to issues

	Distractor Level		Model Type		Question Type
	Word/phrase	Sentence	Extractive	Generative	
(Gao et al., 2019)	Y	Y		Y	R.C.
(Araki et al., 2016)	Y		Y		Cloze
(Guo et al., 2016)	Y		Y		Cloze
(Kumar et al., 2015)	Y	Y	Y		Cloze
(Liang et al., 2017)	Y			Y	Cloze
(Liang et al., 2018)	Y	Y	Y		R.C.
(Chung et al., 2020)		Y		Y	R.C.
(Ren and Q. Zhu, 2021)	Y			Y	Cloze
(Peng et al., 2022)		Y		Y	R.C.
(Chiang et al., 2022)	Y			Y	Cloze
(Zou et al., 2022)		Y	Y	Y	True-False MCQ
this work		Y		Y	True-False MCQ

Table 1: An Overview of the Existing Distractor Generation Methods

such as inconsistent reviewing criteria or unfair judgment. In our paper, we propose the RAD (Retrieval-based Accuracy Differential) metric as an alternative approach for performance evaluation. For clarity of comparison, we summarize the existing DG studies in Table 1.

3 Methodology

Our framework begins with a user-provided keyword, related to a specific topic of interest. As shown in Figure 3, our framework works as follows.

1. **Sentence Retrieval:** From a datastore of learning material, sentences are selected based on their similarity to a given set of keywords.
2. **Keyword-based Sentence Modification:** Using the selected sentences, keywords are chosen and replaced using masked language modeling to generate modified versions of the original sentences.
3. **Sentence Elongation with Autoregressive Models:** Shorter sentences are elongated using autoregressive models to provide continuation for the masked language models during keyword replacement.
4. **Fact Verification:** Modified sentences are passed through a fact verification model to ensure they result in factual inaccuracies, so they can be used as distractors in the questions.
5. **Ranking Using an NLI Premise Model:** Generated sentences are ranked using an NLI premise model, which poses each sentence as

a premise and constructs a hypothesis from a target topic. The ranking is based on the probability of their entailment with the hypothesis.

3.1 Support Sentence Retrieval

We assume a data store consisting of learning material (e.g. the content from a textbook) is available. The first step is to select sentences from the data store and use the sentences as the basis for TF statement generation in the following stage. Specifically, this stage works as follows.

- $D = \{S_1, S_2, \dots, S_N\}$: The datastore consisting of N sentences, where S_i represents the i^{th} sentence.
- K : The given keyword set for sentence retrieval.
- $V(S)$: A function that converts a sentence S into a vector in a vector space.
- $V(K)$: The vector representation of the keyword set K .
- $\text{similarity}(A, B)$: The similarity function between vectors A and B .

The similarity score between the keyword K and a sentence S_i in the datastore can be calculated as:

$$\text{Score}(S_i, K) = \text{similarity}(V(S_i), V(K))$$

To retrieve the top- M sentences from the datastore based on their similarity to the keyword, we calculate the similarity scores for all sentences and select the M sentences with the highest scores:

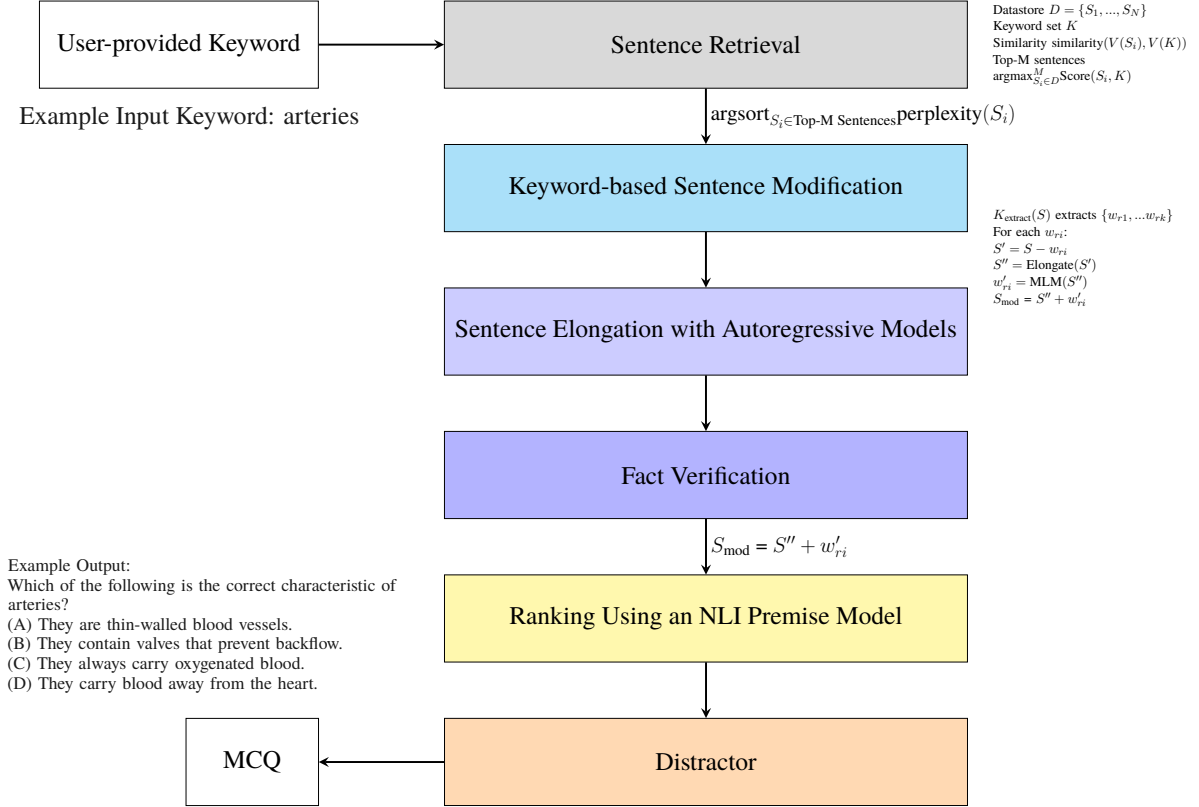


Figure 3: TFDG Process Flow

$$\text{Top-M Sentences} = \text{argmax}_{S_i \in D}^M \text{Score}(S_i, K)$$

This results in a set of sentences from the dataset that are most similar to the given keywords.

Once we have retrieved the top- M sentences, we can further rank them based on their perplexity. Lower perplexity indicates a higher probability and, hence, a better quality or more "expected" sentence. The ranking can be defined as:

$$\text{Ranked Sentences} = \text{argsort}_{S_i \in \text{Top-M Sentences}} \text{perplexity}(S_i)$$

Here, argsort returns the indices that would sort an array, and in this case, it returns the sentences sorted by their perplexity in ascending order. A simplified example of this process is provided in Table 5 in the Appendix.

3.2 Keyword Extraction, Sentence Elongation, and Statement Modification

Once the sentences are retrieved, the subsequent phase in our TFDG pipeline encompasses the extraction of pivotal keywords from these sentences. These extracted keywords are foundational in altering the original sentences to formulate diverse True-False statement options.

- $K_{\text{extract}}(S)$: A function to extract the top- k keywords from a sentence S . This results in a ranked list of keywords $\{w_{r1}, w_{r2}, \dots, w_{rk}\}$. In the implementation of this study, we use KeyBERT model (Giarelis et al., 2021) for the keyword extraction purposes.
- S' : The sentence after masking a selected keyword.
- S'' : The elongated version of S' produced using an autoregressive language model. In this study, we use GPT3 (Floridi and Chiriatti, 2020) for sentence elongation.
- w'_{ri} : The word suggested by the MLM (Masked Language Modeling) to replace the masked keyword w_{ri} in S' . In this study, we also use GPT3 (Floridi and Chiriatti, 2020) for MLM token generation.

For every keyword w_{ri} extracted from a given sentence:

1. Mask the keyword w_{ri} in the sentence, producing S' .
2. Prior to employing the MLM, utilize an autoregressive model to elongate S' , resulting in

S'' . This step is driven by the observation that shorter sentences often lack detailed context, making it challenging for MLMs to produce specific or apt predictions.

3. With S'' as input, invoke a Masked Language Model to suggest a replacement w'_{r_i} for the masked keyword.
4. Integrate w'_{r_i} back into the original sentence to generate a plausible false statement.

By utilizing a keyword extraction process, combined with sentence elongation, the method ensures that significant terms are recognized and appropriately manipulated. The elongated context provided by the autoregressive model facilitates the MLM in making more contextually relevant replacements. This process is illustrated in Table 6, which presents a simplified example of keyword-based sentence modification. Table 7 further demonstrates the application of sentence elongation with autoregressive models.

This methodology offers a systematic avenue to morph sentences retrieved from data stores into potential True-False question candidates. Ensuing stages in the pipeline will delve into framing these as cohesive questions and affirming their educational relevance, as shown in Table 8, which provides a simplified example of statement modification.

3.3 Fact Verification for Statement Validation

After generating modified sentences, it is vital to ascertain that these sentences are indeed false or incorrect. This step is crucial when creating single-choice questions, as having multiple correct answers can introduce ambiguity and confuse the test-takers. To tackle this challenge, we employ a fact verification model.

- S_{mod} : The modified sentence post keyword replacement.
- $FV(S)$: A fact verification function that outputs ‘True’ if statement S is factually accurate, and ‘False’ otherwise. In this study, we use Chatgpt for this purpose.

The verification process can be outlined as:

1. Input the modified sentence S_{mod} into the fact verification function FV .

2. If $FV(S_{\text{mod}})$ returns ‘True’, this suggests that the modification did not alter the factual correctness of the sentence. In such cases, additional modifications or alternative strategies should be considered.
3. If $FV(S_{\text{mod}})$ returns ‘False’, it confirms that the modified sentence is factually incorrect and can be utilized as a distractor in TF MCQ questions.

By integrating the fact verification model, we ensure that the modified statements are genuinely incorrect, thereby preserving the integrity and reliability of the single-choice questions. A simplified example of the fact verification process is illustrated in Table 9.

3.4 Ranking Using an NLI Premise Model

Once the sentences have been generated and verified for factual inaccuracy, we proceed to rank them based on their relevance and quality with the help of a Natural Language Inference (NLI) premise model. The idea is to understand the intrinsic meaning and intent behind each sentence and compare it to a target topic or concept.

- S_{gen} : A sentence generated in the prior stage.
- K : Target topic keywords, e.g., ”arteries”.
- $H(S, K)$: A function that constructs a hypothesis based on sentence S_{gen} and topic K . For instance, given S_{gen} and $K = \text{”arteries”}$, the hypothesis might be ”The sentence S_{gen} is about arteries”.
- $P_{\text{entailment}}(S, H)$: The probability that sentence S_{gen} entails the hypothesis H .

The ranking process involves:

1. For each generated sentence S_{gen} , construct a hypothesis $H(S_{\text{gen}}, K)$ based on the target topic K .
2. Input S_{gen} and $H(S_{\text{gen}}, K)$ into the NLI model to get the entailment probability $P_{\text{entailment}}(S_{\text{gen}}, H)$.
3. Rank the sentences based on the obtained entailment probabilities. A higher probability indicates that the sentence is more relevant and of better quality concerning the indicated topic.

Question Set	Accuracy		RAD
	Without Reference	With Reference	
Basic TCE Questions	0.52	0.60	+0.08
Advanced TCE Questions	0.42	0.37	-0.05
English crackSAT.net Questions	0.59	0.62	+0.03

Table 2: Validity Verification of the RAD Metric

By leveraging the NLI premise model, we can filter out sentences that do not align closely with the desired topic, ensuring that only the most pertinent and high-quality sentences are selected. In the implementation, we use mDeBERTa-v3-base (Yin et al., 2019) as the NLI model. A simplified example of this ranking process using the NLI premise model is shown in Table 10.

4 Evaluation

4.1 RAD Validation

4.1.1 RAD Implementation

As previously discussed, we introduced the RAD metric as a means to gauge the effectiveness of our framework. A well-crafted MCQ should effectively distinguish between students familiar with the material and those who are not, embodying high discriminative power. To validate this, every generated MCQ underwent two separate evaluations. In the first evaluation, ChatGPT was solely presented with the MCQ to determine an answer. In the subsequent evaluation, additional relevant text was integrated into the MCQ, procured using a retrieval method. This direct comparison—highlighted by the difference in the model’s accuracy—serves as a metric for assessing an MCQ’s discriminative power. A greater difference indicates enhanced discriminative capability. To retrieve text associated with each MCQ, the KeyBert model was employed to extract three key terms from every MCQ option. Using these 12 keywords, 12 relevant sentences were retrieved with Pyserini (Lin et al., 2021) from our testing corpus. These sentences were then concatenated and incorporated into the prompts for MCQ answering.

4.1.2 RAD Validation Result

To validate the efficacy of the RAD metric, we applied it to real examination questions to determine whether a significant RAD value could be observed in questions created by human teachers. For this purpose, we selected true/false type multiple-choice questions from two question banks for Biology:

- the Taiwan College Entrance (TCE) Examination question bank, available at <https://testbank.hle.com.tw/>
- SAT Biology questions from CrackSAT.net, accessible at <https://www.cracksat.net/>

The TCE biology question bank is divided into two categories: basic questions and advanced questions. It contains 50 basic questions, 100 advanced questions from the TCE exam, and 47 questions from CrackSAT.net. These questions, curated by the examination center, were designed by expert educators to assess students’ knowledge and understanding of the subject matter. The rigorous scrutiny they have undergone ensures their quality, making them suitable candidates for validating RAD. We present the results of this experiment in Table 2.

For the basic TCE questions, the model initially showed an accuracy of 0.52. However, after the inclusion of reference material, this accuracy increased to 0.60. This improvement, indicated by a RAD value of +0.08, was observed in the human-designed multiple-choice questions (MCQs). Similar results were noted in the English questions from CrackSAT.net, where accuracy improved from 0.59 to 0.62. An interesting observation was that the model struggled with the complexities of the advanced TCE questions, achieving an accuracy of only 0.42. Intriguingly, the introduction of reference materials appeared to have a negative impact, with accuracy decreasing to 0.37. We hypothesize that the reason for this could be that more difficult questions often require logical reasoning beyond mere rote memorization. The presence of additional reference information might have introduced distractions and noise, impeding the model’s ability to answer correctly.

4.2 Results on the Discriminative Power of TFDG as Indicated by the RAD Metric

4.2.1 Corpus and Keywords for TFDG

Our evaluation of the TFDG framework’s performance leveraged the RAD metric. The experiment utilized two authoritative sources for sentence retrieval and subsequently applied the RAD metric to evaluate the outcomes: the specialized Biology textbook for the Taiwan College Entrance (TCE) Examination (https://www.hle.com.tw/book_detail/?code=HBI1-1) and AP

Data Sets	Accuracy		RAD
	Without Reference	With Reference	
TCE Biology	0.50	0.68	+0.18
SAT Biology	0.36	0.47	+0.11

Table 3: TFDG’s RAD Result

courses from OpenStax (ISBN-13: 978-1-947172-41-8) and Barron’s for SAT Biology (eISBN: 978-1-4380-6812-1). The keywords for inputting TFDG were extracted from basic TCE questions and English crackSAT.net questions.

- **TCE Biology Dataset:** An increase in RAD value of +0.18, from an accuracy of 0.50 without reference material to 0.68 with it, indicates that the TFDG framework has a discriminative capacity when enriched with contextual content from the Taiwan College Entrance examination’s Biology textbook. This suggests that the framework is highly effective in differentiating between students’ knowledge states.
- **SAT Biology Dataset:** For the SAT Biology dataset, an increase in accuracy from 0.36 to 0.47 and a corresponding RAD value of +0.11 also reflect the TFDG’s discriminative effectiveness, albeit to a lesser extent compared to the TCE dataset. The rise in the RAD value here demonstrates that the TFDG framework can ensure the discriminative power of the generated MCQs.

The experimental results, as presented in Table 3, showcase the TFDG framework’s ability to discern the depth of a student’s understanding. The RAD metric’s role in this experiment was pivotal, offering a quantifiable measure of the improvement in the MCQs’ ability to discriminate based on the availability of reference information. Through this, the TFDG framework’s potential in creating nuanced and educationally valuable MCQs that can effectively test a student’s grasp of the subject matter is confirmed.

4.3 Ablation Study

To dissect the inner workings of the TFDG framework, we embarked on an ablation study, assessing the impact of individual components on the performance across two different datasets: TCE Biology and SAT Biology. The TFDG framework was evaluated in its full form and in two variant conditions where specific components were omitted:

Dataset	Condition	Accuracy		RAD
		Without Reference	With Reference	
TCE	Full	0.50	0.68	+0.18
	w/o FV	0.38	0.58	+0.20
	w/o Elongation	0.35	0.73	+0.38
SAT	Full	0.36	0.47	+0.11
	w/o FV	0.28	0.40	+0.12
	w/o Elongation	0.27	0.49	+0.22

Table 4: Ablation Evaluation of TCE and SAT Biology Datasets

Fact Validation (FV) and Elongation. Specifically, the experimental setup included three variants of the TFDG pipeline: (1) **[Full]**: The full TFDG framework, (2) **[w/o FV]**: TFDG without Fact Validation, and (3) **[w/o Elongation]**: TFDG without Elongation.

The results, summarized in Table 4, reveal insights into our design.

- **Impact of Fact Validation (FV):** Without FV, accuracy decreases in the ‘without reference’ condition due to multiple correct answer options generated by TFDG, causing confusion. However, adding references significantly improves accuracy, suggesting references help resolve uncertainties caused by the absence of FV.
- **Elongation’s Role in Clarity:** The ‘w/o Elongation’ condition demonstrates lower accuracy without references, emphasizing Elongation’s importance in generating clear options. With references, accuracy improves, indicating references help address ambiguities arising from the lack of Elongation.
- **Efficacy of the Full TFDG Framework:** The Full TFDG condition, including FV and Elongation, starts with higher baseline accuracy without references, indicating clear questions with a single correct non-factual statement. Adding references doesn’t substantially improve accuracy, suggesting FV and Elongation enhance the quality of generated MCQs by introducing ‘confusable’ options.

5 Conclusion

In this paper, we address two main issues: how to automatically create incorrect True-False options and how to assess the quality of these generated options. Specifically, we propose a pipeline that generates True-False incorrect options based on user-provided keywords. Additionally, we introduce the RAD metric to evaluate the generated results. Preliminary experiments demonstrate that

our pipeline effectively generates medium-level questions, as evidenced by the RAD metric comparison. However, our current architecture struggles to generate more challenging questions that require reasoning and logical judgment. Therefore, our current achievements are primarily applicable to modifying literal distractors. Furthermore, we also need to refine the RAD metric to account for cases where the initial model’s answer accuracy is low due to multiple correct options in the generated results.

6 Limitations

The advantage of this architecture is its ability to automatically generate multiple-choice questions for any preprocessed text. It can be applied to various competency tests or assist teachers in generating multiple-choice questions related to specific domains in the field of education.

But our architecture only focuses on processing and replacing the text content within the articles, which imposes limitations on its applications. If the text requires reasoning and logical thinking, the performance of TFDG framework may not meet expectations, such as in the case of mathematics or philosophy-related content. Additionally, this architecture is unable to generate more diverse multiple-choice questions and can only provide True/False type questions.

In the field of education, the principle of teaching according to the student’s ability is highly significant. While our framework might be capable of generating questions based on topics that students are less proficient in, it lacks the capability to adjust the difficulty level according to individual students’ proficiency. This presents a potential direction for future research.

Acknowledgement

This work is supported by NSTC 112-2634-F-005 -002-project Smart Sustainable New Agriculture Research Center (SMARTer), NSTC Taiwan Project under grant 112-2221-E-005 -075 -MY3, and Delta Research Center, Delta Electronics, Inc.

References

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th*

International Conference on Computational Linguistics: Technical Papers, pages 1125–1136.

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

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

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6423–6430.

Nikolaos Giarelis, Nikos Kanakaris, and Nikos Karacapilidis. 2021. A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17*, pages 635–645. Springer.

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: Generating knowledge assessments for arbitrary topics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI’16)*. AAAI Press.

Girish Kumar, Rafael E Banchs, and Luis Fernando D’Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161.

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

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

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, pages 2356–2362.

Hsien-Yung Peng, Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2022. Misleading inference generation via proximal policy optimization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 497–509. Springer.

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

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

Chak Yan Yeung, John SY Lee, and Benjamin K Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Han Cheng Yu, Yu An Shih, Kin Man Law, KaiYu Hsieh, Yu Chen Cheng, Hsin Chih Ho, Zih An Lin, Wen-Chuan Hsu, and Yao-Chung 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, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2019. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension.

Bowei Zou, Pengfei Li, Liangming Pan, and Aiti Aw. 2022. Automatic true/false question generation for educational purpose. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 61–70.

Appendix

Support Sentence Retrieval (TCE Example)

Input: 生態系 (En: ecosystem)

Retrieved Results:

- **1986 年，科學家提出「生物多樣性」一詞，早期此名詞使用於生態研究時清查該地區的所有生物種類，並以「物種數」表示。(En: In 1986, scientists proposed the term ‘biodiversity.’ In early ecological research, this term was used to inventory all biological species in a given area, represented by the ‘number of species.’)**
- 外來入侵種易對各類原生物種產生危害，對當地物種多樣性造成衝擊。(En: Invasive alien species easily harm various native species and impact local biodiversity.)
- 河流下游多為沙洲泥地，水生植物是水中消費者的養分來源。(En: The downstream river areas are often sandbars and mudflats, where aquatic plants serve as a nutrient source for aquatic consumers.)
- 遠洋區位於近海區之外，水深超過 200 公尺，平均可達 4000 公尺。(En: The pelagic zone is located beyond the coastal zone, with depths exceeding 200 meters and an average depth reaching 4000 meters.)

Support Sentence Retrieval (SAT Example)

Input: PLANT FORM AND PHYSIOLOGY

Retrieved Results:

- Mammals use uric acid crystals as an antioxidant in their cells.
- An organ system is a higher level of organization that consists of functionally related organs.
- Mammalian sex determination is determined genetically by the presence of X and Y chromosomes.
- **The periderm substitutes for the epidermis in mature woody-stemmed plants .**

Table 5: Simplified Example for Support Sentence Retrieval. Note that as introduced in Sec. 3.1, we will retrieve the Top-M sentences. In the example shown above, the sentences highlighted in bold will be used in the following Table to complete the entire pipeline and form the distractor options, while the remaining sentences will be used as general options in the final step (Table 10).

Keyword-based Sentence Modification (TCE Example)

Input: 1986 年，科學家提出「生物多樣性」一詞，早期此名詞使用於生態研究時清查該地區的所有生物種類，並以「物種數」表示。(En: In 1986, scientists proposed the term ‘biodiversity.’ In early ecological research, this term was used to inventory all biological species in a given area, represented by the ‘number of species.’)

Output: “生物多樣性 (En: biodiversity)”, “生態研究 (En: ecological research)”, “物種數 (En: number of species)”

Post-processing:

- 1986 年，科學家提出「[MASK]」一詞... (En: In 1986, scientists proposed the term ‘[MASK].’...)
- ... 早期此名詞使用於 [MASK] 時清查該地區的所有生物種類... (En: ... In early [MASK], this...)
- ... 並以「[MASK]」表示。(En: ... area, represented by the ‘[MASK].’)

Keyword-based Sentence Modification (SAT Example)

Input: The periderm substitutes for the epidermis in mature woody-stemmed plants.

Output: ”periderm”, ”epidermis”, ”mature woody-stemmed plants”

Post-processing:

- The [MASK] substitutes for the epidermis in mature woody-stemmed plants.
- The periderm substitutes for the [MASK] in mature woody-stemmed plants.
- The periderm substitutes for the epidermis in [MASK].

Table 6: Simplified Example for Keyword-based Sentence Modification

Sentence Elongation with Autoregressive Models (TCE Example)

Input: 1986 年，科學家提出「生物多樣性」一詞，早期此名詞使用於生態研究時清查該地區的所有生物種類，並以「物種數」表示。(En: In 1986, scientists proposed the term 'biodiversity.' In early ecological research, this term was used to inventory all biological species in a given area, represented by the 'number of species.')

Output: 2000 年左右，隨著人類活動的減少，生物多樣性之數量逐漸消退，生物多樣性之數量隨之萎縮。生物多樣性中心於 2001 年對人類活動頻繁的地區進行數位化調查。(En: Around the year 2000, as human activities decreased, the quantity of biodiversity gradually declined, and the number of species shrank. In 2001, the Biodiversity Center conducted a digital survey in areas with frequent human activity.)

Sentence Elongation with Autoregressive Models (SAT Example)

Input: The periderm substitutes for the epidermis in mature woody-stemmed plants.

Output: However, unlike the epidermis, the periderm is less effective in protecting against water loss and disease. It consists mainly of cork cells, which can sometimes become too thick and rigid, impairing gas exchange and making the plant more vulnerable to environmental stress.

Table 7: Simplified Example for Sentence Elongation with Autoregressive Models

Statement Modification (TCE Example)

Input: 1986 年... 早期此名詞使用於 [MASK] 時清查該地區的所有生物種類... 並以「物種數」表示。2000 年左右... 數位化調查。 (En: In 1986... In early [MASK], this... by the 'number of species.' **Around the year 2000,... with frequent human activity.**)

Output: “生物學 (En: Biology)”, “生態系統 (En: Ecosystem)”, “環境保護 (En: Environmental Protection)”, “生態平衡 (En: Ecological Balance)”, “自然資源 (En: Natural Resources)”, “生態群落 (En: Ecological Community)”, “動物學 (En: Zoology)”, “生物演化 (En: Biological Evolution)”, “生物地理學 (En: Biogeography)”, “生態保育 (En: Ecological Conservation)”

Post-processing:

- ... 早期此名詞使用於**生物學**時清查該地區的所有生物種類... (En: ... In early **biology**, this...)
- ... 早期此名詞使用於**生態系統**時清查該地區的所有生物種類... (En: ... In early **ecosystem**, this...)
- :
- ... 早期此名詞使用於**生態保育**時清查該地區的所有生物種類... (En: ... In early **ecological conservation**, this...)

Statement Modification (SAT Example)

Input: The periderm substitutes for the epidermis in [MASK]. **However, unlike the epidermis, the periderm.....to environmental stress.**

Output: ”ferns”, ”grasses”, ”herbs”, ”aquatic plants ”, ”mosses”, ”cacti”, ”lichens”, ”annuals”, ”algae”, ”succulents”

Post-processing:

- The periderm substitutes for the epidermis in **ferns**.
- The periderm substitutes for the epidermis in **grasses**.
- :
- The periderm substitutes for the epidermis in **succulents**.

Table 8: Simplified Example for Statement Modification. Note that in the Keyword-based Sentence Modification step (Table 6), there are multiple results, and we only use one as a demonstration, placed in the first half of the input in the example above. In reality, each result goes through this step. The bold text in the second half corresponds to the output from Table 7, which will be directly appended to enhance the effect of statement modification.

Fact Verification (TCE Example)

Input: ... 早期此名詞使用於**生物學**時清查該地區的所有生物種類... (En: ... In early **biology**, this...)

Output: True

Input: ... 早期此名詞使用於**生態系統**時清查該地區的所有生物種類... (En: ... In early **ecosystem**, this...)

Output: True

Input: ... 早期此名詞使用於**環境保護**時清查該地區的所有生物種類... (En: ... In early **environmental protection**, this...)

Output: False

Input: 1986 年，科學家提出「**物種多樣性**」一詞... (En: In 1986, scientists proposed the term '**species diversity**.'...)

Output: True

Input: 1986 年，科學家提出「**物種分類學**」一詞... (En: In 1986, scientists proposed the term '**species taxonomy**.'...)

Output: False

:
:

Fact Verification (SAT Example)

Input: The periderm substitutes for the epidermis in **cacti**

Output: False

Input: The periderm substitutes for the epidermis in **succulents**

Output: False

Input: The periderm substitutes for the **bark** in mature woody-stemmed plants.

Output: False

Input: The periderm substitutes for the **pith** in mature woody-stemmed plants.

Output: False

Input: The periderm substitutes for the epidermis in **ferns**

Output: False

:
:

Table 9: Simplified Example for Fact Verification

Ranking Using an NLI Premise Model (TCE Example)

Input: ... 早期此名詞使用於**環境保護**時清查該地區的所有生物種類... (En: ... In early **environmental protection**, this...)

Score: 0.806

Input: ... 早期此名詞使用於**生態群落**時清查該地區的所有生物種類... (En: ... In early **ecological community**, this...)

Score: 0.457

Input: 1986 年，科學家提出「**物種分類學**」一詞... (En: In 1986, scientists proposed the term '**species taxonomy**.'...)

Score: **0.823 (highest)**

Post-processing (Generating MCQ):

Which of the following statements is wrong?

(A) 1986 年，科學家提出「**物種分類學**」一詞... (En: In 1986, scientists proposed the term '**species taxonomy**.'...)

(B) 外來入侵種易對各類原生物種產生危害... (En: Invasive alien species easily...)

(C) 河流下游多為沙洲泥地... (En: The downstream river areas are...)

(D) 遠洋區位於近海區之外... (En: The pelagic zone is located...)

Ans: (A)

The correct statement should be: 1986 年，科學家提出「**生物多樣性**」一詞... (En: In 1986, scientists proposed the term '**biodiversity**.'...)

Ranking Using an NLI Premise Model (SAT Example)

Input: The periderm substitutes for the **sclerenchyma** in mature woody-stemmed plants.

Score: **0.796 (highest)**

Input: The periderm substitutes for the **bark** in mature woody-stemmed plants.

Score: 0.221

Input: The periderm substitutes for the sclerenchyma in **herbs**.

Score: 0.521

Post-processing (Generating MCQ):

Which of the following statements is wrong?

(A) Mammals use uric acid crystals as an antioxidant in their cells.

(B) An organ system is a higher level of organization that consists of functionally related organs.

(C) Mammalian sex determination is determined genetically by the presence of X and Y chromosomes.

(D) The periderm substitutes for the **sclerenchyma** in mature woody-stemmed plants.

Ans: (D)

The correct statement should be: The periderm substitutes for the epidermis in mature woody-stemmed plants.

Table 10: Simplified Example for Ranking Using an NLI Premise Model. Note that in the example shown above, the options other than the distractors (such as options (B), (C), and (D) in the TCE Example, and options (A), (B), and (C) in the SAT Example) are sentences retrieved in the Support Sentence Retrieval step (Table 5).