Chain-of-Exemplar: Enhancing Distractor Generation for Multimodal Educational Question Generation

Haohao Luo1, Yang Deng2, Ying Shen1,3,4*, See-Kiong Ng2, Tat-Seng Chua2
1Sun Yat-sen University  2National University of Singapore  3Pazhou Lab  4Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology

Sheny76@mail.sysu.edu.cn {ydeng,seekiong,dcscst}@nus.edu.sg

1 Introduction

Multiple-choice questions (MCQs) are important in enhancing concept learning and student engagement for educational purposes. Despite the multimodal nature of educational content, current methods focus mainly on text-based inputs and often neglect the integration of visual information. In this work, we study the problem of multimodal educational question generation, which aims at generating subject-specific educational questions with plausible yet incorrect distractors based on multimodal educational content. To tackle this problem, we introduce a novel framework, named Chain-of-Exemplar (CoE), which utilizes multimodal large language models (MLLMs) with Chain-of-Thought reasoning to improve the generation of challenging distractors. Furthermore, CoE leverages three-stage contextualized exemplar retrieval to retrieve exemplary questions as guides for generating more subject-specific educational questions. Experimental results on the ScienceQA benchmark demonstrate the superiority of CoE in both question generation and distractor generation over existing methods across various subjects and educational levels.

Abstract

Multiple-choice questions (MCQs) are important in enhancing concept learning and student engagement for educational purposes. Despite the multimodal nature of educational content, current methods focus mainly on text-based inputs and often neglect the integration of visual information. In this work, we study the problem of multimodal educational question generation, which aims at generating subject-specific educational questions with plausible yet incorrect distractors based on multimodal educational content. To tackle this problem, we introduce a novel framework, named Chain-of-Exemplar (CoE), which utilizes multimodal large language models (MLLMs) with Chain-of-Thought reasoning to improve the generation of challenging distractors. Furthermore, CoE leverages three-stage contextualized exemplar retrieval to retrieve exemplary questions as guides for generating more subject-specific educational questions. Experimental results on the ScienceQA benchmark demonstrate the superiority of CoE in both question generation and distractor generation over existing methods across various subjects and educational levels.

1 Introduction

Multiple-choice questions (MCQs) are important in education for promoting deep and extensive knowledge acquisition. Research (Davis, 2009) indicates that well-crafted questions with educational intent are closely linked to heightened student engagement and achievement. A key aspect in question generation is the quality of the distractors (Gierl et al., 2017). Questions with inadequate or low-quality distractors are less challenging and easier to solve. Generating plausible, yet incorrect distractors is crucial in educational contexts, as effective and challenging distractors can significantly enhance students’ reading comprehension and contribute to their overall academic success. However, it is costly and time-consuming to manually produce MCQs, since even professional test developers do not manage to write more than three or four good MCQs per day (Kim et al., 2012).

To alleviate the human labour, automatic MCQ generation has received extensive attention. Previous research (Berre et al., 2022; Liang et al., 2018; Ren and Zhu, 2021; Qiu et al., 2020) primarily focuses on text-based inputs for MCQ generation, while generating MCQs from multimodal contexts is still relatively underexplored. Such emphasis on text often leads to underutilization of visual information, which is prevalent within educational content, such as textbooks (Lu et al., 2022) or examinations (Zhang et al., 2023a). Additionally, some latest studies (Wang and Baraniuk, 2023) neglect the creation of challenging and thought-provoking distractors, which is essential for high-quality educational question generation. Moreover, questions generated by current methods tend to be too general and not tailored for specific subjects or educational levels. As illustrated in Figure 1, the question generation model is tasked with creating educational questions from a biology textbook that includes both textual descriptions and visual illustrations. The example question focuses on subject-specific knowledge, and the accompanying distractors are carefully crafted to be plausible yet incorrect, enhancing the educational value of the questions.

In the light of these challenges, we propose a novel framework named Chain-of-Exemplar (CoE), which combines retrieved exemplars and Chain-of-Thought (CoT) reasoning to generate educational

Figure 1: Illustration of our multimodal educational question and distractor generation problem.
questions and distractors from multimodal inputs of texts and images. Specifically, we employ multimodal large language models (MLLMs) to encode multimodal contexts and incorporate them into a three-stage multimodal-CoT framework that separates question generation, rationale generation, and distractor generation. The CoT helps trigger the reasoning capability of MLLMs leading to the generation of plausible and confusing distractors. Meanwhile, to generate more specialized questions for educational purposes in specific subjects, we leverage retrieved exemplary educational questions as demonstrations to guide the generation. Finally, we adopt an easy-to-apply multi-task training strategy to finetune our generative models.

The main contributions of this work can be summarized as follows:

- We propose a three-stage framework, namely CoE, to generate customized questions and plausible distractors, via multi-task finetuning MLLMs to perform multimodal-CoT reasoning.
- To enhance the generation of specialized educational questions in specific subjects, we utilize retrieved exemplary educational question as demonstration to guide the generation.
- Experimental results on ScienceQA benchmark show that CoE outperforms existing methods and effectively takes advantage of MLLMs. Our code will be released via https://github.com/Luohh5/Chain-of-Exemplar.

2 Related Works

Question Generation Question generation (QG) (Pan et al., 2019) plays a crucial role in applications like conversational systems (Gao et al., 2019; Do et al., 2023; Zeng et al., 2023; Deng et al., 2022) and intelligent tutoring systems (Xu et al., 2022; Yao et al., 2022; Zhang et al., 2022; Zhao et al., 2022; Dugan et al., 2022; Deng et al., 2023b). Evolving from prior research based on syntactic trees or knowledge bases (Heilman and Smith, 2010; Kumar et al., 2015), most existing studies typically adopt deep neural networks (Du et al., 2017; Li et al., 2019; Dong et al., 2024) for question generation. With the advent of pre-trained and large language models (PLMs/LLMs), recent works (Bulathwela et al., 2023; Wang et al., 2022; Wang and Baraniuk, 2023) design various fine-tuning strategies to enhance the QG capabilities of language models. Regarding educational purposes, multiple-choice question generation (Berre et al., 2022) holds great importance, where distractor generation (Ren and Zhu, 2021; Qiu et al., 2020) plays a crucial role. Apart from text-based inputs, there is also a growing body of research focusing on QG from images (Mostafazadeh et al., 2016; Li et al., 2018). However, it is worth noting that most existing work focuses on uni-modal QG, leaving the potential of multimodal QG largely unexplored.

Multimodal Question Answering Since question generation serves as the inverse task of question answering (QA), addressing the challenges in this field effectively necessitates drawing insights from QA studies. Due to the multimodal nature of information flow in real-world applications, researchers (Hannan et al., 2020; Talmor et al., 2021; Luo et al., 2023) emphasize the importance of answering questions that require information across multiple modalities, which is typically referred as multimodal question answering. Notably, several studies have focused on multimodal question answering in educational contexts, such as textbook-based (Lu et al., 2022) and exam-based (Zhang et al., 2023a) questions. Generating questions based on these contexts holds great potential for constructing intelligent tutoring systems and facilitating personalized learning experiences for students.

Chain-of-Thought Reasoning Recently, to solve complex reasoning tasks, CoT prompting (Wei et al., 2022) is proposed to decompose complex problems into a series of intermediate steps by prompting LLMs. Subsequently, the CoT reasoning has been effectively applied in various contexts, including multi-modal reasoning (Zhang et al., 2023b; Wu et al., 2023), multi-lingual scenarios (Shi et al., 2023; Qin et al., 2023), dialogue systems (Wang et al., 2023a; Deng et al., 2023a), and knowledge-driven applications (Trivedi et al., 2023; Wang et al., 2023b). Apart from these, there has been a surge in the development of other Chain-of-X methods and most of them primarily focus on augmenting the LLMs with guidance to improve reasoning capabilities. For instance, Chain-of-Knowledge (Li et al., 2023) augments LLMs by dynamically incorporating grounding information from heterogeneous sources for more factual rationales. Chain-of-Note (Yu et al., 2023) augments LLMs with a series of reading notes to retrieve
documents for more precise and contextualized reasoning. Inspired by these works, we equip LLMs with retrieved exemplars for contextual knowledge supplementation and guidance for generation.

3 Method

Given a correct answer $A$ and multimodal context $C = \{I, T\}$, where $I$ represents the image and $T$ represents the textual paragraph, the task of multimodal educational question generation aims to generate a relevant question $Q$ and several distracting answers $A'$. The overview of the proposed CoE framework is illustrated in Figure 2.

3.1 Model Architecture

As illustrated in Figure 2, CoE is composed of four distinct modules: a question generator module $G_{QG}$, a rationale generator module $G_{RG}$, a distractor generator module $G_{DG}$, and a Contextualized Exemplar Retrieval (CER) module $R$. The question generator, rationale generator, and distractor generator utilize the same pre-trained multimodal language models (e.g., Qwen-VL) as the backbone with sharing weights. By employing these three generators, we introduce a CoT reasoning strategy to decompose multi-step problems into intermediate reasoning steps (rationale) and then generate the distractors. To guide the generation, we introduce a similar Contextualized Exemplar Retrieval module (Section 3.2) to retrieve the most relevant example from training data and use it as demonstration for a given test instance.

3.2 Contextualized Exemplar Retrieval

In order to retrieve a similar sample as exemplar $E$, for more subject-specific generation, we introduce a Contextualized Exemplar Retrieval (CER) module to discern the analogy between each sample in training data $D$ and associate them, as shown in Figure 2. Specifically, we first encode the attribute information (i.e., textual context $T$, answer $A$, and question $Q$) of each example into a vector using the Angle (Li and Li, 2023).

$$\mathcal{Y}_t = \mathcal{M}(T), \mathcal{Y}_a = \mathcal{M}(A), \mathcal{Y}_q = \mathcal{M}(Q),$$

where $\mathcal{M}(-)$ and $\mathcal{V}$ denote the encoder and vector representations, respectively. All the vectors lie in a latent sample space that contains rich semantics. If two vectors are close in the latent space, they are more likely to share similar information in analogous field. Subsequently, we calculate the cosine similarity of each attribute vector between the given testing instance $S$ and each other samples $S^i \in D$, then retrieve the nearest neighbor in the latent space as the most relevant example:

$$I = \arg \max_{i \in \{1, 2, ..., N\}} \max (Sim^i_t, Sim^i_a, Sim^i_q),$$

$$Sim^i_t = \frac{(\mathcal{V}^i_t)^T \mathcal{V}_t}{\|\mathcal{V}^i_t\|_2 \|\mathcal{V}_t\|_2},$$

$$Sim^i_a = \frac{(\mathcal{V}^i_a)^T \mathcal{V}_a}{\|\mathcal{V}^i_a\|_2 \|\mathcal{V}_a\|_2},$$

$$Sim^i_q = \frac{(\mathcal{V}^i_q)^T \mathcal{V}_q}{\|\mathcal{V}^i_q\|_2 \|\mathcal{V}_q\|_2},$$

where $I$ denotes an index of the most similar sample among all $N$ samples and $E = S^I$. We concatenate the test instance with the retrieved exemplar into a prompt for formatted input and feed it into the generator modules:

$$X_{QG} = \{A, T, I, E_{QG}\},$$

$$X_{RG} = \{Q, A, T, I, E_{RG}\},$$

$$X_{DG} = \{Q, R, A, T, I, E_{DG}\},$$

where $E_{QG}$ contains exemplar image, context, answer and question while $E_{RG}$ and $E_{DG}$ are further
expanded with rationale and distractors. In the subsequent generation process, the retrieved exemplar provides supplementary contextual knowledge that may not be present in test instance’s context and exerts flexible control of the output to make its style similar to the exemplar, which is especially effective for example with limited context. In this manner, the CER module retrieves relevant information as supplementary for the original sample to ground the generation on the subject at hand.

3.3 Chain-of-Exemplar Reasoning

To construct the framework of Chain-of-Exemplar, we combine the CER module and Chain-of-Thought (CoT) reasoning to generate educational questions and distractors. Specifically, The CoE reasoning framework consists of three generation stages: (i) question generation, (ii) rationale generation, and (iii) distractor generation. All three stages share the same model architecture but differ in the input and output formats.

**Question Generation** In the question generation stage, we feed the question generator with retrieved exemplar $E_{QG}$, answer input $A$, and context input $C$ including textual paragraph $T$ and associated image $I$. The primary objective is to train a question generation model $G_{QG}$:

$$Q = G_{QG}(A, T, I, E_{QG}).$$

**Rationale Generation** In the rationale generation stage, the generated question $Q$ is appended to the original input $X_{QG} = \{A, T, I, E_{QG}\}$ and the exemplar $E_{QG}$ is supplemented with corresponding rationale as $E_{RG}$ to construct the further input in the second stage, $X_{RG} = \{Q, A, T, I, E_{RG}\}$.

Then, we feed the updated input to the rationale generation model to generate intermediate reasoning as the rationale.

$$R = G_{RG}(Q, A, T, I, E_{RG}).$$

**Distractor Generation** Similarly, the input in the final distractor generation stage is constructed by expanding the exemplar $E_{RG}$ with corresponding distractors and concatenating the generated rationale $R$ with the previous input $X_{RG}$ as $X_{DG} = \{Q, R, A, T, I, E_{DG}\}$. Subsequently, we feed the modified input to the distractor generator by

$$A' = G_{DG}(Q, R, A, T, I, E_{DG}),$$

where $A'$ denotes the plausible yet incorrect answers for the question $Q$.

3.4 Multi-task Training Procedure

After formatting all prompt inputs, we perform instruction fine-tuning on a multimodal large language model in a multi-task way. Specifically, we assemble the formatted data by combining and shuffling all examples from the three tasks: question generation, rationale generation, and distractor generation. Following the teacher forcing method, we utilize the groundtruth question and rationale as input in distractor generation. Then we minimize the sum of negative log-likelihood loss $L_{NLL}$ averaged over tokens in three generation tasks as our training objective:

$$L_{NLL} = -\frac{1}{L} \sum_{l=1}^{L} \bar{y}_{i} \log \left( \frac{\exp (y_{i})}{\sum_{l} \exp (y_{l})} \right),$$

$$L_{total} = L_{NLL}^{QG} + L_{NLL}^{RG} + L_{NLL}^{DG},$$

where $L$ is the max length of output sequence, $\bar{y}_{i}$ and $y_{i}$ denote the $i$-th token in the groundtruth sequence and prediction sequence respectively. By training the MLLMs on these tasks simultaneously, our goal is to prevent intermediate errors during CoT training that may disrupt reasoning, as well as induce the model being more robust to the wording choices of the prompts.

3.5 Inference

The inference phase also consists of question generation, rationale generation, and distractors generation stages. Given the image $I$, context $T$, answer $A$, and retrieved exemplar $E_{QG}$, the question generator generates corresponding questions $Q$ for next stage. Subsequently, the rationale generator utilizes all the aforementioned inputs along with the generated question and expanded exemplar $E_{RG}$ to generate rationales $R$ for intermediate reasoning. Finally, the distractor generator use all the previous inputs, including the generated rationale and augmented exemplar $E_{DG}$, to predict plausible distractors $A'$. It’s worth noting that we only calculate the maximum between answer and context similarity for exemplar retrieval since the question for the test instance is not given during inference.

4 Experiment

4.1 Experimental Settings

**Datasets** We conduct the experiments on the reversed ScienceQA dataset (Lu et al., 2022). ScienceQA is the first large-scale multimodal educational dataset that annotates detailed lectures and
Table 1: Dataset statistics of ScienceQA benchmark. Question types: NAT = natural science, SOC = social science, LAN = language science, TXT = containing text context, IMG = containing image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Modality</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>NAT</td>
<td>SOC</td>
</tr>
<tr>
<td>#</td>
<td>11,487</td>
<td>4,350</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics of ScienceQA benchmark. Question types: NAT = natural science, SOC = social science, LAN = language science, TXT = containing text context, IMG = containing image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

4.2 Evaluation on Question Generation

We first conduct evaluation on the question generation, including both automatic evaluation and human evaluation.

4.2.1 Automatic Evaluation

Table 2 presents the comparison of the automatic evaluation results of CoE with previous state-of-the-art models, which demonstrates that all the strong baselines fail to compete with CoE on both BLEU4 and ROUGE-L. Among the baselines, MultiQG-TI and Multimodal-CoT, which instantiate the question generator with multimodal large language models, largely outperform VL-T5 which simply extends pre-trained language models with visual understanding ability. Meanwhile, compared with MultiQG-TI that leverages image captions in the context to provide vision semantics, Multimodal-COT achieves much better performance by utilizing image features. Furthermore, the results clearly show that ChatGPT fails at the multimodal QG task in our setting. Although its performance steadily improves with more examples in the in-context learning setting, ChatGPT trails CoE by a significant margin.

Additionally, among the 3 subject classes, all the question generation baselines consistently demonstrate superior performance in social science (SOC) while exhibiting the lowest performance in language science (LAN). They also achieve performance gain for the questions with paired images (IMG), but perform poorly in the absence of any textual or image hints (NO). Moreover, the performance of CoE exhibits high consistency across different subjects and grades, which justifies the generalizability of the framework in education field.

However, the BLEU-4 and ROUGE-L metrics solely focus on evaluating the exact match between the generated question and the groundtruth, neglecting the aspect of question diversity. To address this concern, we incorporate an additional metric to evaluate the question diversity automatically and report it in Appendix B.

7982
### 4.2.2 Human Evaluation

We further conduct human evaluation to evaluate the quality of the generated questions and investigate if the generated distractors can confuse the examinees in the real human test. For question generation, we randomly select 50 question samples generated by different methods and employ three annotators with good English background to rate them from 1 (worst) to 5 (best) based on 4 metrics: (1) **Readability** measures whether the generated questions are easy to read and understand for students of corresponding grade; (2) ** Appropriateness** examines whether the generated questions are aligned with the corresponding subjects; (3) **Complexity** estimates the level of reasoning or cognitive effort required to answer the generated question for students of corresponding grade; (4) **Engagement** measures whether the students find the questions engaging and have interest in answering the questions. The annotator guideline is presented in Appendix E.

Table 3 illustrates the human evaluation results of question generation. Although the groundtruth questions win the highest score on all the metrics, our proposed **CoE** outperforms all the rest baselines and achieves a very close performance to the groundtruth. Furthermore, an average score of 4.48 indicates that our model can reliably generate challenging and thought-provoking educational questions that exhibit impressive readability, appropriateness, complexity, and engagement. Notably, the usage of MLLMs brings huge results difference between Multimodal-CoT and VL-T5. Interestingly, although ChatGPT performs poorly on most metrics, it demonstrates superior performance in readability and engagement in comparison to VL-T5, which indicates its strong few-shot learning capability of generating readable paragraphs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject</th>
<th>Modality</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B-4/↑/R-L↑)</td>
<td>NAT</td>
<td>SOC</td>
<td>LAN</td>
</tr>
<tr>
<td>ChatGPT 0 shot</td>
<td>1.2/18.4</td>
<td>3.8/25.6</td>
<td>0.1/10.1</td>
</tr>
<tr>
<td>ChatGPT 1 shot</td>
<td>4.5/22.3</td>
<td>2.1/22.1</td>
<td>1.3/14.1</td>
</tr>
<tr>
<td>ChatGPT 3 shot</td>
<td>5.6/22.4</td>
<td>8.5/26.5</td>
<td>6.5/22.1</td>
</tr>
<tr>
<td>VL-T5</td>
<td>55.7/72.4</td>
<td>71.8/80.0</td>
<td>34.9/49.5</td>
</tr>
<tr>
<td>MultiQG-TI</td>
<td>63.2/80.7</td>
<td>80.3/88.3</td>
<td>40.9/55.8</td>
</tr>
<tr>
<td>Multimodal-CoT</td>
<td>76.1/84.7</td>
<td>85.4/91.2</td>
<td>59.4/68.6</td>
</tr>
<tr>
<td>CoE</td>
<td>83.2/89.1</td>
<td>93.2/95.6</td>
<td>66.1/72.8</td>
</tr>
</tbody>
</table>

Table 2: Automatic evaluation results of question generation. ↑: higher is better, ↓: lower is better.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT</td>
<td>4.47</td>
<td>2.77</td>
<td>2.18</td>
<td>2.90</td>
<td>3.08</td>
</tr>
<tr>
<td>VL-T5</td>
<td>3.41</td>
<td>3.58</td>
<td>3.69</td>
<td>2.33</td>
<td>3.26</td>
</tr>
<tr>
<td>MultiQG-TI</td>
<td>4.55</td>
<td>4.29</td>
<td>3.99</td>
<td>4.30</td>
<td>4.28</td>
</tr>
<tr>
<td>Multimodal-CoT</td>
<td>4.55</td>
<td>4.41</td>
<td>4.09</td>
<td>4.35</td>
<td>4.35</td>
</tr>
<tr>
<td>CoE</td>
<td>4.65</td>
<td>4.58</td>
<td>4.29</td>
<td>4.39</td>
<td>4.48</td>
</tr>
</tbody>
</table>

Table 3: Human evaluation results of question generation. ↑: higher is better, ↓: lower is better.

### 4.3 Evaluation on Distractor Generation

We then evaluate the performance in terms of distractor generation for constructing multiple-choice questions, also including both automatic evaluation and human evaluation.

#### 4.3.1 Automatic Evaluation

Table 4 summarizes the automatic evaluation results for distractor generation. Similar to QG, MultiQG-TI and Multimodal-CoT demonstrate superior performance in comparison to VL-T5 due to the strong cognition and generation capabilities of MLLMs. Notably, giving the credit to the leveraging of CoT reasoning, the distractors generated by Multimodal-CoT are more confusing and challenging than MultiQG-TI. Similarly, there remains a significant disparity between the performance of ChatGPT and CoE.

In further analysis of the results across 3 subjects, we observe that the performance of all evaluated methods does not exhibit consistent superiority within a specific subject, which is different from QG. In fact, our CoE demonstrates its highest performance in the domain of natural science (NAT), while yielding the poorest performance in social science (SOC), showcasing a notable deviation from QG outcomes. Besides, CoE achieves impressive results in terms of the ROUGE-L score for questions accompanied by paired text (TXT), albeit with compromised accuracy. Conversely,
questions lacking both paired text and image (NO) exhibit a reduction in ROUGE-L score, without impacting accuracy, which highlights the strong adaptability in no contexts distractor generation of CoE. Similarly, the high performance consistency across different subjects and grades of our CoE framework provides further evidence of its generalizability.

4.3.2 Human Evaluation

Moreover, we invited another three annotators to answer the selected 50 MCQs with the generated distractors from different methods as well. The Accuracy of their answering would serve as a human evaluation metric for assessing the quality of generated distractors. Besides, we also adopt a 5-point scale for other 3 metrics to evaluate the quality of generated distractors including: (1) Overlap examines whether the generated distractors are completely overlapping with the correct answer; (2) Plausibility estimates whether the generated distractors are semantically relevant to the given context and question; (3) Distinctiveness measures the originality of the generated distractors in comparison to the groundtruth. Elaborate evaluation guideline is depicted in Appendix E.

Table 5 summarizes the human evaluation results of distractor generation. During the ablation study, the intermediate rationale could indeed enhance the distractor generation.

Table 4: Automatic evaluation results of distractor generation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject</th>
<th>Modality</th>
<th>Grade</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAT</td>
<td>SOC</td>
<td>LAN</td>
<td>TXT</td>
</tr>
<tr>
<td>ChatGPT 0 shot</td>
<td>81.7/29.7</td>
<td>90.8/41.2</td>
<td>71.9/24.3</td>
<td>80.8/32.8</td>
</tr>
<tr>
<td>ChatGPT 1 shot</td>
<td>81.9/42.8</td>
<td>84.6/46.5</td>
<td>73.3/33.8</td>
<td>82.7/42.5</td>
</tr>
<tr>
<td>ChatGPT 3 shot</td>
<td>82.8/45.9</td>
<td>85.8/47.2</td>
<td>76.7/38.6</td>
<td>84.4/42.3</td>
</tr>
<tr>
<td>VL-T5</td>
<td>80.3/58.0</td>
<td>76.4/50.5</td>
<td>74.4/36.7</td>
<td>85.5/60.7</td>
</tr>
<tr>
<td>MultiQG-TI</td>
<td>78.6/61.2</td>
<td>74.6/53.4</td>
<td>73.7/39.3</td>
<td>83.7/63.8</td>
</tr>
<tr>
<td>Multimodal-CoT</td>
<td>68.2/65.6</td>
<td>78.1/53.9</td>
<td>70.2/56.3</td>
<td>76.4/73.4</td>
</tr>
</tbody>
</table>

| CoE             | 62.4/72.0 | 74.5/57.1 | 64.0/62.0 | 71.2/78.6 | 72.4/55.9 | 61.5/62.3 | 67.0/65.7 | 62.1/68.3 | 65.4/66.6 |

Table 5: Human evaluation results of distractor generation. During the ablation study, the intermediate rationale could indeed enhance the distractor generation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Overl.</th>
<th>Plaus.</th>
<th>Dist.</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT 0 shot</td>
<td>1.37</td>
<td>2.63</td>
<td>4.51</td>
<td>94.67</td>
</tr>
<tr>
<td>VL-T5</td>
<td>3.83</td>
<td>3.66</td>
<td>3.15</td>
<td>92.67</td>
</tr>
<tr>
<td>MultiQG-TI</td>
<td>3.21</td>
<td>4.19</td>
<td>3.34</td>
<td>91.33</td>
</tr>
<tr>
<td>Multimodal-CoT</td>
<td>1.08</td>
<td>4.41</td>
<td>3.39</td>
<td>83.33</td>
</tr>
<tr>
<td>CoE</td>
<td>0.24</td>
<td>4.89</td>
<td>3.61</td>
<td>79.33</td>
</tr>
<tr>
<td>Groundtruth</td>
<td>0.05</td>
<td>4.93</td>
<td>-</td>
<td>87.33</td>
</tr>
</tbody>
</table>

Table 4: Automatic evaluation results of distractor generation.

Table 5: Human evaluation results of distractor generation. During the ablation study, the intermediate rationale could indeed enhance the distractor generation.

4.4 Ablation Study

We perform ablation studies to investigate the effects of the proposed approaches in terms of chain-of-thought reasoning, Contextualized Exemplar Retrieval module, and multi-task learning, as presented in Table 6. There are several notable observations as follows:

- When dropping the Chain-of-Thought (CoT) reasoning, the distractors are directly generated based on context and question with associated answer, whose performance drops by 15.7% Acc and 0.9 R-L, respectively. The results show that the intermediate rationale could indeed enhance the distractor generation.

- When we drop the Contextualized Exemplar Retrieval (CER) module, the overall performance declines a lot, especially in terms of a significant -14.3% drop in Acc, which demonstrates that adding the CER module indeed helps retrieve supplemental information for the original sample and benefit generation. Besides, More details about the ablation study of CER module are presented in Appendix C.
Table 6: Ablation study on question and distractor generation. ↑: higher is better, ↓: lower is better.

- When we use single-task learning in place of multi-task learning (MTL) as our finetune strategy, we can see declines in both generation tasks, specifically a decrease of -6.4 in B-4 score and an increase of +13.1% in Acc score, which further verifies the effectiveness of utilizing multi-task learning to mitigate intermediate errors during CoT finetuning.

- As for the input context, dropping either the image or the textual paragraph results in a substantial decline in performance for both generation tasks, particularly with a significant -14.8 and -21.0 decrease in B-4. It highlights the advantages of incorporating both visual and textual information when generating questions and distractors.

4.5 Case Study
To qualitatively evaluate the four modules in CoE, we visualize an example from ScienceQA in Figure 2. Based on the context, image, and answer input in this example, the question generator is obviously powerless to generate appropriate question without referring to the exemplar question "Which property do these three objects have in common?". Moreover, the exemplar rationale "An object has different properties ... property that all three objects have in common is salty" provides valuable information supplementary for the rationale generator and serves as a demonstration to guide the generation. Furthermore, generating such diverse and plausible distractors is benefit from the informative intermediate reasoning which explains the definition of "property" and describes the commonalities of the objects in image.

To further estimate how the CER module and CoT reasoning affect the question and distractor generation, we present generated examples in Figure 3 and Figure 4. The example in Figure 3 illustrates that when we add the CER module, the generated question "Which property do these four objects have in common?" which asks about the "object property" exhibits more impressive complexity and appropriateness for education. Conversely, when dropping the exemplar, the generated question "What do the tastes of these four foods in images like?" appears simplistic and trivial for answering. Moreover, Figure 4 demonstrates that when we drop out the intermediate reasoning, the distractor generator fails to reason from the explanation regarding "what kind of animal’s feet have grabbing prey adaptation", which results in generating distinct distractor "bighorn sheep" and irrelevant distractor "octopus".

5 Conclusions
In this paper, we present a novel framework called Chain-of-Exemplar (CoE), which combines retrieved exemplars and Chain-of-Thought (CoT) reasoning to generate educational questions and distractors from multimodal inputs of texts and images. Specifically, we utilize MLLMs to encode multimodal contexts and incorporate them into a three-stage multimodal-CoT framework, namely question generation, rationale generation, and distractor generation. Meanwhile, we introduce a Contextualized Exemplar Retrieval (CER) module to retrieve exemplary educational questions as demonstrations to guide the generation. We finally adopt an easy-to-apply multi-task training strategy to finetune our generative models. Our experiments on ScienceQA benchmark demonstrate that CoE outperforms existing methods and achieves new state-of-the-art performance.
Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative, and NExT Research Center. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

Limitations

Distinctiveness of Generation  As mentioned in Appendix B, our proposed CoE can achieve a much better performance than most baselines in question generation. Meanwhile, we acknowledge the limitation of the question diversity compared to ChatGPT. Similar in Section 4.3.2, while the distractor generation of CoE exhibits a superior distinctiveness compared to the majority of baselines, it still falls short of meeting annotators’ expectations. In contrast to our CoE, which employ supervised learning and the diversity of generation heavily relies on the finetuning data, ChatGPT is able to generate highly distinctive question and distractors by utilizing in-context learning without any supervision. Consequently, there remains room for future research to explore effective finetuning strategies and investigate how to incorporate external knowledge or distill knowledge from ChatGPT into the open-source LLMs.

Hallucination Issues  Another limitation of CoE is that our rationale generation module may suffer from the typical flaw of hallucination issues, i.e., making fabricated intermediate reasoning that is irrelevant to the context and answer. Hallucinated rationale would mislead the generation process, resulting in irrelevant and trivial distractors generation. One potential solution involves replacing the greedy decoding strategy used in Chain-of-Thought reasoning with self-consistency (Wang et al., 2023c) by sampling diverse reasoning paths and selecting the most consistent rationale for distractor generation, which mimic multiple different ways of thinking. The greater the diversity of reasoning paths, the more authentic and reasonable rationale can be generated by CoT. We believe that this future research direction will prove valuable and promising in effectively tackling hallucination issues.

Exemplar Resource During training, our CER module retrieves domain-specific exemplars from training split of ScienceQA to guide the generation. However, the generators trained under the guidance of these exemplars greatly restricts the quality and diversity of the generated questions and distractors, primarily due to the strong dependence on training data. Actually, this limitation is not exclusive to our work. The educational MCQ generation heavily relies on training data in specific domain to generate questions and distractors for educational purpose. Therefore, we acknowledge the need for future research to explore methods for incorporating the retrieved exemplar with external knowledge to reduce the reliance on training data and enhance the generation quality.

References


Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yuya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. CoRR, abs/2304.14178.


Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a.
M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. 
CoRR, abs/2306.05179.


Zhao, Zhenjie; Hou, Yufang; Wang, Dakuo; Yu, Mo; Liu, Chengzhong; Ma, Xiaojuan. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. 

A Implementation Details

We use Qwen-7B (Bai et al., 2023) as the backbone for question, rationale, and distractor generation. Notably, we reproduce Multimodal-CoT (Zhang et al., 2023b) by still employing Qwen-7B in place of Flan-T5-Large (Chung et al., 2022) as the backbone. For training, we set the max length of both input and output sequence to 2048. Due to insufficient CUDA memory, we utilize Q-LoRA (Dettmers et al., 2023) as our finetune strategy and reduce the batch size to 2. Besides, we finetune the model up to 5 epochs, with a maximum learning rate of $1e^{-5}$, a minimum learning rate of $1e^{-6}$, and a linear warmup of 3000 steps. Additionally, we employ the language encoder of pretrained AnglE-LLaMA-7B (Li and Li, 2023) as the backbone for CER module. Our experiments are run on 4 NVIDIA GTX 3090 24G GPUs. The prompting details of CoT are presented in Figure 5.

For ChatGPT baselines, we utilize the gpt-3.5-turbo-1106 model API throughout ChatGPT zero-shot and few-shots experiments. The prompts we give to ChatGPT, which are almost the same as CoE prompts in both question and distractor generation, are shown below.

```markdown
## 0-shot Question Generation
Context: ... Answer: ...
Generate a question based on the corresponding context and answer.

## 1-shot Question Generation
Context: ... Answer: ...
Refer to the example, generate a question based on the corresponding context and answer. Exemplar: ...

## 2-shot Question Generation
Context: ... Answer: ...
Refer to these 2 examples, generate a question based on the corresponding context and answer. Exemplar 1: ... Exemplar 2: ...

## 3-shot Question Generation
Context: ... Answer: ...
Refer to these 3 examples, generate a question based on the corresponding context and answer. Exemplar 1: ... Exemplar 2: ... Exemplar 3: ...

## 0-shot Distractor Generation
Context: ... Answer: ...
Based on the above context and answer, generate at least 1 plausible yet incorrect answers and separate them with numbers like (1) (2) (3).

## 1-shot Distractor Generation
Context: ... Answer: ...
Refer to the example and based on the above context and answer, generate at least 1 plausible yet incorrect answers and separate them with numbers like (1) (2) (3). Exemplar: ...

## 2-shot Distractor Generation
Context: ... Answer: ...
Refer to these 2 examples and based on the above context and answer, generate at least 1 plausible yet incorrect answers and separate them with numbers like (1) (2) (3). Exemplar 1: ... Exemplar 2: ...

## 3-shot Distractor Generation
Context: ... Answer: ...
Refer to these 3 examples and based on the above context and answer, generate at least 1 plausible yet incorrect answers and separate them with numbers like (1) (2) (3). Exemplar 1: ... Exemplar 2: ... Exemplar 3: ...
```
Table 7: Distinct-n scores of question generation.

<table>
<thead>
<tr>
<th>Retrieval Strategy</th>
<th>Question Generation B-4 ↑</th>
<th>R-L ↑</th>
<th>Acc ↓</th>
<th>R-L ↑</th>
<th>Distractor Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>70.4</td>
<td>79.3</td>
<td>78.6</td>
<td>58.3</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>81.1</td>
<td>86.5</td>
<td>65.4</td>
<td>66.6</td>
<td></td>
</tr>
<tr>
<td>Summation</td>
<td>81.6</td>
<td>86.9</td>
<td>65.5</td>
<td>67.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Detailed performance in terms of different exemplar retrieval strategies. ↑: higher is better, ↓: lower is better.

C Analysis of Exemplar Retrieval

Analysis of Exemplar Retrieval Strategies We construct an experiment by training the generators with different exemplar retrieval strategies to investigate whether the exemplar retrieval strategy influences the performance. Specifically, we utilize 3 retrieval strategies: Random, Maximum, and Summation. In "Random" strategy, exemplars are randomly selected from the training data. The "Maximum" strategy denotes using the argmax for the maximum between answer, context, and question similarities, while the "Summation" strategy denotes combining the three signals by summing up them as follows:

$$I = \arg \max_{i \in \{1, 2, \ldots, N\}} \left( Sim^i_v + Sim^i_a + Sim^i_q \right),$$

$$Sim^i_v = \frac{(V^i_v)^T V^i_v}{\|V^i_v\|_2 \|V^i_v\|_2},$$

$$Sim^i_a = \frac{(V^i_a)^T V^i_a}{\|V^i_a\|_2 \|V^i_a\|_2},$$

$$Sim^i_q = \frac{(V^i_q)^T V^i_q}{\|V^i_q\|_2 \|V^i_q\|_2},$$

As depicted in Table 8, the performance of randomly selected exemplars fails to compete with that of contextually retrieved exemplars in both question and distractor generation, which further justify that contextually retrieved exemplar indeed provides valuable and useful information to the generators and validate the effective of contextually retrieval. Moreover, retrieving exemplar using "Summation" strategy demonstrates superior performance in comparison to "Maximum" strategy, indicating that combining the three signals in exemplar retrieval yields more appropriate and relevant exemplars that benefit both question and distractor generation.

B Question Diversity

To measure the diversity of generated questions, we utilize Distinct-n scores (Li et al., 2016) as an automatic evaluation metric. Specifically, it calculates the number of distinct n-grams in corpus-level, with a higher count indicating a greater diversity of questions. We consider values of n ranging from 1 to 4. As presented in Table 7, the performance of all methods improves as the value of n increases from 1 to 4. Additionally, ChatGPT demonstrates the capability to generate highly distinctive questions by utilizing zero-shot or few-shot in-context learning without any supervision, which is similar to the performance in distractor generation. Except to ChatGPT, our CoE outperforms all other baselines, which indicates that the questions generated by CoE exhibit impressive appropriateness while maintaining high diversity.
Effect of the Number of Exemplars  
To further analyse the potential impact of the number of exemplars on generation performance, we vary the number of exemplars, denoted as $N$, and retrieve the top $N$ exemplars with the highest similarity. As shown in Figure 6, we observe an improvement in both question and distractor generation performance as $N$ increases from 1 to 3, verifying the effectiveness and utility of information the exemplar provides to the generators. However, we note that when $N = 2$, the limitation of the maximum input length results in the truncation of certain content within the exemplar, preventing it from performing at its optimum capacity. Therefore, both the question and distractor generation performance improve slowly when $N \geq 2$.

D  Analysis of Different Base Models
To analyse the generality of our CoE framework, we conduct an experiment to utilize other base models in place of Qwen-VL (Bai et al., 2023) as the backbone for question and distractor generation, including LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), mPLUG-Owl (Ye et al., 2023), and VisualGLM-6B (Ding et al., 2021; Du et al., 2022). Note that we employ the same prompt for all base models to ensure fairness in the comparison. As summarized in Table 9, Qwen-VL outperforms all the rest base models, showcasing its high applicability and suitability in our framework. Generally, while there are slight difference in performance among the 5 base models, they consistently demonstrate superior performance in both question and distractor generation, which further confirms the effectiveness and versatility of our CoE framework.

Table 9: Detailed performance of our CoE framework with different base models. ↑: higher is better, ↓: lower is better.

<table>
<thead>
<tr>
<th>Method</th>
<th>Question Generation</th>
<th>Distractor Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-4 ↑</td>
<td>R-L ↑</td>
</tr>
<tr>
<td>Qwen-VL</td>
<td>81.1</td>
<td>86.5</td>
</tr>
<tr>
<td>LLaVA</td>
<td>80.7</td>
<td>86.0</td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>80.5</td>
<td>86.0</td>
</tr>
<tr>
<td>mPLUG-Owl</td>
<td>80.3</td>
<td>85.8</td>
</tr>
<tr>
<td>VisualGLM-6B</td>
<td>51.2</td>
<td>73.2</td>
</tr>
<tr>
<td></td>
<td>Acc ↓</td>
<td>R-L ↑</td>
</tr>
<tr>
<td>Qwen-VL</td>
<td>65.4</td>
<td>66.6</td>
</tr>
<tr>
<td>LLaVA</td>
<td>70.1</td>
<td>60.3</td>
</tr>
<tr>
<td>InstructBLIP</td>
<td>69.7</td>
<td>60.0</td>
</tr>
<tr>
<td>mPLUG-Owl</td>
<td>72.3</td>
<td>58.0</td>
</tr>
<tr>
<td>VisualGLM-6B</td>
<td>77.5</td>
<td>39.6</td>
</tr>
</tbody>
</table>

E  Guideline of Generation Quality Evaluation
We present the guideline of human evaluation for question and distractor generation quality in Figure 7.
Guideline of Generation Quality Evaluation

This study aims to evaluate the quality of the question and distractor generation. Each case provides you with a context, image, and groundtruth. You need to evaluate the generated question and distractors from the following aspects.

**Case**

**Context:** Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem.

**Answer:** copepod

**Groundtruth Question:** Which of the following organisms is the primary consumer in this food web?

**Groundtruth Distractors:** (1) black crappie (2) bacteria

---

**Question Evaluation**

- **Readability:** whether the generated questions are easy to read and understand for students of corresponding grade.

  - **Options:**
    1. Complete understanding
    2. Quite understanding
    3. Moderate understanding
    4. Minor understanding
    5. No understanding

  - **Examples:**
    1. “Which of the following organisms is the primary consumer in this food web?” exhibits a high level of readability and can be completely understood.
    2. “Which organism is the primary consumer in this food web?” demonstrates moderate readability since it has some typos and grammatical mistakes.
    3. “How is the matter eaten by organisms?” has no linguistic logic and shows no readability, resulting in difficulty to understand.

- ** Appropriateness:** whether the generated questions are semantically aligned with the corresponding subjects.

  - **Options:**
    1. Completely appropriate
    2. Mostly appropriate
    3. Fairly appropriate
    4. Mostly inappropriate
    5. Completely inappropriate

  - **Examples:**
    1. This case belongs to natural science.
    2. “Which of the following organisms is the primary consumer in this food web?” shows fairly appropriateness in natural science domain.

- **Complexity:** the level of reasoning or cognitive effort required to answer the generated question for students of corresponding grade.

  - **Options:**
    1. Completely relevant
    2. Mostly relevant
    3. Fairly relevant
    4. Slightly relevant
    5. Not relevant

  - **Examples:**
    1. “Which organism preys on golden algae in this food web?” is fairly relevant to answer.
    2. “What is the largest fish in this picture?” is completely irrelevant to natural science.

- **Engagement:** whether the students find the questions engaging and have interest in answering the questions.

  - **Options:**
    1. Completely interesting
    2. Mostly interesting
    3. Fairly interesting
    4. Mostly uninteresting
    5. Completely uninteresting

  - **Examples:**
    1. “Which organism preys on golden algae in this food web?” is completely irrelevant to natural science domain.

---

**Distractor Evaluation**

- **Consistency:** whether the generated distractors are completely overlapping with the correct answer.

  - **Options:**
    1. Completely consistent
    2. Mostly consistent
    3. Moderately consistent
    4. Mostly inconsistent
    5. Completely inconsistent

  - **Examples:**
    1. “copepod” is completely overlapping with the correct answer.
    2. “Copepoda” shows consistentency with the correct answer.

- **Plausibility:** whether the generated distractors are semantically relevant to the given context and question.

  - **Options:**
    1. Completely relevant
    2. Mostly relevant
    3. Moderately relevant
    4. Mostly irrelevant
    5. Completely irrelevant

  - **Examples:**
    1. “(1) kelp (2) zooplankton” are both organisms in fresh water lake but not in Little Rock Lake. Therefore, they are moderately relevant to the given context.
    2. “(1) sandpiper (2) falcon” are both flying organisms, which are completely irrelevant to the context.

- **Distinctiveness:** the originality of the generated distractors in comparison to the groundtruth.

  - **Options:**
    1. Completely distinctive
    2. Mostly distinctive
    3. Moderately distinctive
    4. A little distinctive
    5. Fully overlapping

  - **Examples:**
    1. “(1) algae (2) bacteria” exhibit a moderate level of distinctiveness since one of them differs from the groundtruth distractors.
    2. “(1) rotifer (2) water flea” are completely distinct from the groundtruth distractors.

---

**Figure 7:** Guideline of human evaluation for question and distractor generation quality.