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

Haohao Luo[1], **Yang Deng**[2], **Ying Shen**[1,3,4*], **See-Kiong Ng**[2], **Tat-Seng Chua**[2]
[1]Sun Yat-sen University [2]National University of Singapore [3]Pazhou Lab
[4]Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology
luohh5@mail2.sysu.edu.cn {ydeng,seekiong,dcscts}@nus.edu.sg
sheny76@mail.sysu.edu.cn

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,

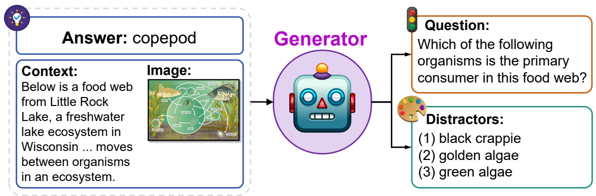


Figure 1: Illustration of our multimodal educational question and distractor generation problem.

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

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7978–7993
August 11-16, 2024 ©2024 Association for Computational Linguistics

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 finetuning 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

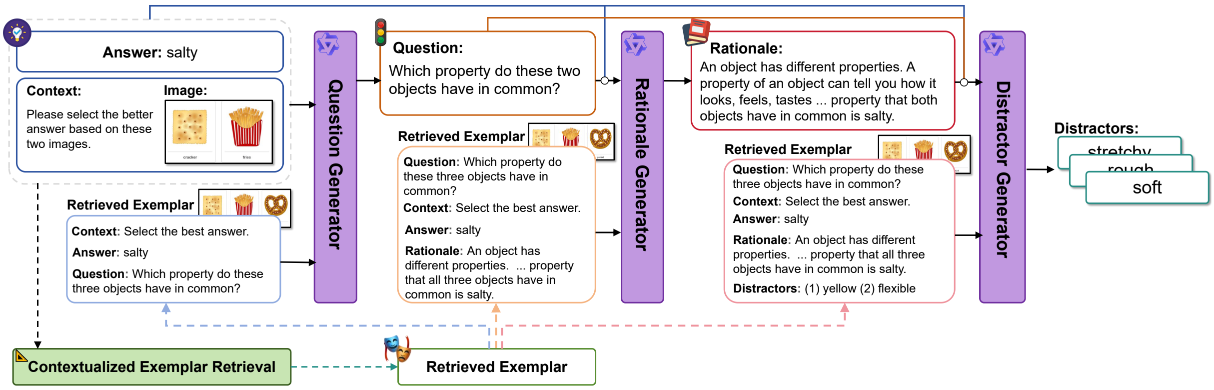


Figure 2: The overall framework of **CoE**.

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 texual 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 $\mathcal{G}_{QG}$, a rationale generator module $\mathcal{G}_{RG}$, a distractor generator module $\mathcal{G}_{DG}$, and a Contextualized Exemplar Retrieval (CER) module $\mathcal{R}$. 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 $\mathcal{E}$ for more subject-specific generation, we introduce a Contextualized Exemplar Retrieval (CER) module to discern the analogy between each sample in training data $\mathcal{D}$ 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).

$$\mathcal{V}_t = \mathcal{M}(T),\ \mathcal{V}_a = \mathcal{M}(A),\ \mathcal{V}_q = \mathcal{M}(Q), \quad (1)$$

where $\mathcal{M}(\cdot)$ 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 $S^i \in \mathcal{D}$, then retrieve the nearest neighbor in the latent space as the most relevant example:

$$\begin{aligned}
\mathcal{I} &= \underset{i\in\{1,2\ldots,N\}}{\arg\max}\ \max(Sim_t^i, Sim_a^i, Sim_q^i),\\
Sim_t^i &= \frac{(\mathcal{V}_t^i)^T\mathcal{V}_t}{\|\mathcal{V}_t^i\|_2\,\|\mathcal{V}_t\|_2},\\
Sim_a^i &= \frac{(\mathcal{V}_a^i)^T\mathcal{V}_a}{\|\mathcal{V}_a^i\|_2\,\|\mathcal{V}_a\|_2},\\
Sim_q^i &= \frac{(\mathcal{V}_q^i)^T\mathcal{V}_q}{\|\mathcal{V}_q^i\|_2\,\|\mathcal{V}_q\|_2},
\end{aligned} \quad (2)$$

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

$$X_{QG} = \{A, T, I, \mathcal{E}_{\mathcal{QG}}\}, \quad (3)$$
$$X_{RG} = \{Q, A, T, I, \mathcal{E}_{\mathcal{RG}}\}, \quad (4)$$
$$X_{DG} = \{Q, R, A, T, I, \mathcal{E}_{\mathcal{DG}}\}, \quad (5)$$

where $\mathcal{E}_{QG}$ contains exemplar image, context, answer and question while $\mathcal{E}_{\mathcal{RG}}$ and $\mathcal{E}_{\mathcal{DG}}$ 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 $\mathcal{E}_{QG}$, 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 $\mathcal{G}_{QG}$:

$$Q = \mathcal{G}_{QG}(A, T, I, \mathcal{E}_{QG}). \tag{6}$$

Rationale Generation In the rationale generation stage, the generated question Q is appended to the original input $X_{QG} = \{A, T, I, \mathcal{E}_{QG}\}$ and the exemplar $\mathcal{E}_{QG}$ is supplemented with corresponding rationale as $\mathcal{E}_{RG}$ to construct the further input in the second stage, $X_{RG} = \{Q, A, T, I, \mathcal{E}_{RG}\}$. Then, we feed the updated input to the rational generation model to generate intermediate reasoning as the rationale.

$$R = \mathcal{G}_{RG}(Q, A, T, I, \mathcal{E}_{RG}), \tag{7}$$

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

$$A' = \mathcal{G}_{DG}(Q, R, A, T, I, \mathcal{E}_{DG}), \tag{8}$$

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 $\mathcal{L}_{NLL}$ averaged over tokens in three generation tasks as our training objective:

$$\mathcal{L}_{NLL} = -\frac{1}{L}\sum_{l=1}^{L} \overline{y_l} \log \left(\frac{\exp(y_l)}{\sum_i^l \exp(y_i)} \right), \tag{9}$$

$$\mathcal{L}_{total} = \mathcal{L}_{NLL}^{QG} + \mathcal{L}_{NLL}^{RG} + \mathcal{L}_{NLL}^{DG}, \tag{10}$$

where L is the max length of output sequence, $\overline{y_l}$ and y_l denote the l-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 $\mathcal{E}_{QG}$, 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 $\mathcal{E}_{RG}$ to generate rationales R for intermediate reasoning. Finally, the distractor generator use all the previous inputs, including the generated rationale and augmented exemplar $\mathcal{E}_{DG}$, 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 Setups

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

	Subject			Modality			Grade	
Type	NAT	SOC	LAN	IMG	TXT	NO	G1-6	G7-12
#	11,487	4,350	5,371	10,332	10,220	7,188	15,422	5,786

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.

explanations for the answers. It contains 21,208 multimodal science questions with rich domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills, showing outstanding generalizability across different domains. The benchmark dataset is split into train, validation, and test sets with 12,726, 4,241, and 4,241 examples, respectively. Note that in order to transform the format from QA to QG setup, we reverse the ScienceQA data by utilizing the context and correct answer as input, and generating the corresponding multi-choice question as output. The details of dataset statistics are presented in Table 1.

Evaluation Metrics We adopt both automatic and human evaluation to measure the performance of our method. Specifically, we choose 2 automatic evaluation metrics including **BLEU-4** (Papineni et al., 2002) and **ROUGE-L** (Lin, 2004) for question generation, both of which have been widely used in existing QG works. Additionally, we choose **ROUGE-L** and **Accuracy** (Chung et al., 2020) as automatic evaluation metrics for distractor generation. Specifically in **Accuracy**, we replace the origin options in ScienceQA with a combination of correct answer and the generated distractors, and leave the rest of the data unchanged. We adopt a multimodal question answering model (Zhang et al., 2023b) (trained by ScienceQA dataset) to evaluate the accuracy of the "modified" multiple-choice questions. Therefore, a higher accuracy score indicates poorer generation quality.

Baselines We compare `CoE` with the state-of-the-art (SOTA) methods in ScienceQA, including VL-T5 (Yeh et al., 2022), MultiQG-Ti (Wang and Baraniuk, 2023), and Multimodal-CoT (Zhang et al., 2023b). Note that we train the Multimodal-CoT by utilizing reversed task format of ScienceQA for question and distractor generation task. Due to limited prior work on automatic multimodal question and distractor generation, we use off-the-shelf model APIs as the baselines. Specifically, we use ChatGPT API (Ouyang et al., 2022) with zero-shot and in-context learning (Kaplan et al., 2020) with up to three examples, each of which is formatted exactly the same as our preprocessed data points in the ScienceQA dataset. More details about ChatGPT baselines and other experimental setups are presented in Appendix A

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.

Method (B-4↑/R-L↑)	Subject			Modality			Grade		Avg
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
ChatGPT 0 shot	1.2/18.4	3.8/25.6	0.1/10.1	1.1/19.7	3.6/25.1	0.1/11.1	1.5/18.2	1.3/17.1	1.4/17.8
ChatGPT 1 shot	4.5/22.3	2.1/22.1	1.3/14.1	4.4/23.6	1.5/21.4	1.2/14.2	2.4/19.7	4.8/21.4	3.2/20.3
ChatGPT 3 shot	5.6/22.4	8.5/26.5	6.5/22.1	6.1/25.2	7.4/25.1	5.4/20.5	6.1/22.9	6.9/23.6	6.4/23.2
VL-T5	55.7/72.4	71.8/80.0	34.9/49.5	53.1/71.5	74.9/82.8	26.5/49.6	53.2/70.7	54.1/67.5	53.7/68.2
MultiQG-TI	63.2/80.7	80.3/88.3	40.9/55.8	61.1/79.8	81.4/90.2	34.0/54.0	61.0/78.7	61.5/71.2	61.2/76.2
Multimodal-CoT	76.1/84.7	85.4/91.2	59.4/68.6	77.0/86.9	86.4/91.5	56.9/68.9	76.4/84.1	68.6/78.0	73.8/82.0
CoE	**83.2/89.1**	**93.2/95.6**	**66.1/72.8**	**84.2/91.2**	**94.6/97.3**	**65.7/73.0**	**83.4/88.7**	**76.6/82.1**	**81.1/86.5**

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

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.

Method	Read.↑	Appro.↑	Com.↑	Eng.↑	Overall↑
ChatGPT	4.47	2.77	2.18	2.90	3.08
VL-T5	3.41	3.58	3.69	2.33	3.26
MultiQG-TI	4.55	4.29	3.99	4.30	4.28
Multimodal-CoT	4.55	4.41	4.09	4.35	4.35
CoE	**4.65**	**4.58**	**4.29**	**4.39**	**4.48**
Groundtruth	4.72	4.92	4.59	4.47	4.68

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,

Method	Subject			Modality			Grade		Avg
(Acc↓/R-L↑)	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
ChatGPT 0 shot	81.7/29.7	90.8/41.2	71.9/24.3	80.8/32.8	88.1/39.5	74.7/27.4	83.1/31.1	77.3/29.8	81.1/30.7
ChatGPT 1 shot	81.9/42.8	84.6/46.5	73.3/33.8	82.7/42.5	81.5/47.5	76.3/37.1	82.8/41.8	75.6/40.6	80.2/41.4
ChatGPT 3 shot	82.8/45.9	85.8/47.2	76.7/38.6	84.4/42.3	83.5/46.7	79.4/42.2	84.9/44.0	76.1/45.2	79.5/44.4
VL-T5	80.3/58.0	76.4/50.5	74.4/36.7	85.5/60.7	73.0/49.8	75.7/37.0	81.7/49.5	71.1/53.9	78.1/51.2
MultiQG-TI	78.6/61.2	74.6/53.4	73.7/39.3	83.7/63.8	73.1/52.3	74.4/41.3	79.4/54.3	70.8/54.5	76.5/54.4
Multimodal-CoT	68.2/65.6	78.1/53.9	70.2/56.3	76.4/73.4	76.9/54.3	68.5/42.5	72.4/60.5	67.2/61.5	70.7/61.0
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 4: Automatic evaluation results of distractor generation. ↑: higher is better, ↓: lower is better.

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 metrics 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. In general, our **CoE** outperforms all other baselines and even surpasses the groundtruth in accuracy, which indicates that the distractors generated by **CoE** are distinctive and thought-provoking enough to distract humans from the correct answer. However, although we have adopted a series of strategies such as CoT and CER module to enhance the plausibility of generated distractors, the distinctiveness remains limited to improve significantly. Contrarily, due to the unknown pretrained knowledge inside the blackbox, ChatGPT often performs well in zero-shot and few-shot text generation, especially in terms of readability and diversity. Therefore, it demonstrates impressive performance in both overlap and distinctiveness but falls short in plausibility, resulting in higher accuracy in annotators' MCQ answering.

Method	Overl.↓	Plaus.↑	Dist.↑	Acc↓
ChatGPT	1.37	2.63	**4.51**	94.67%
VL-T5	3.83	3.66	3.15	92.67%
MultiQG-TI	3.21	4.19	3.34	91.33%
Multimodal-CoT	1.08	4.41	3.39	83.33%
CoE	**0.24**	**4.89**	3.61	**79.33%**
Groundtruth	0.05	4.93	-	87.33%

Table 5: Human evaluation results of distractor generation. ↑: higher is better, ↓: lower is better.

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.

Method	Question Generation		Distractor Generation	
	B-4 ↑	R-L ↑	Acc ↓	R-L ↑
CoE	**81.1**	**86.5**	**65.4**	**66.6**
- w/o MTL	74.7	83.1	78.5	65.4
- w/o CoT	-	-	81.1	65.7
- w/o Exemplars	75.2	82.8	79.7	65.2
- w/o Image	66.3	74.9	81.4	63.1
- w/o Text	60.1	70.2	81.0	64.3

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 with 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*

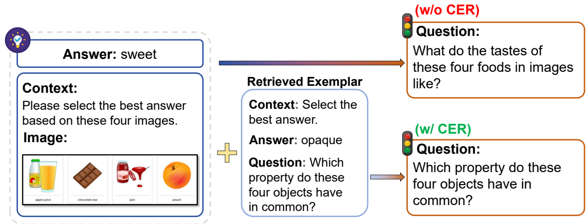


Figure 3: Case study in terms of the CER module.

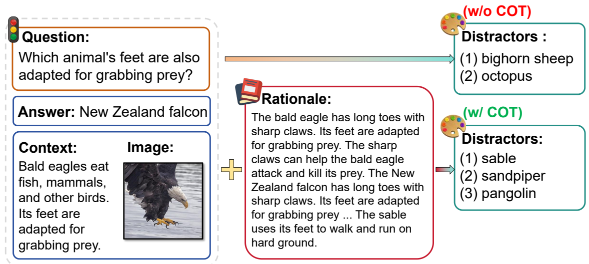


Figure 4: Case study in terms of COT reasoning.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Guillaume Le Berre, Christophe Cerisara, Philippe Langlais, and Guy Lapalme. 2022. Unsupervised multiple-choice question generation for out-of-domain q&a fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 732–738. Association for Computational Linguistics.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *Artificial Intelligence in Education - 24th International Conference, AIED 2023, Tokyo, Japan, July 3-7, 2023, Proceedings*, volume 13916 of *Lecture Notes in Computer Science*, pages 327–339. Springer.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4390–4400. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Barbara Gross Davis. 2009. *Tools for teaching*. John Wiley & Sons.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6970–6984. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10602–10621. Association for Computational Linguistics.

Yang Deng, Zifeng Ren, An Zhang, Wenqiang Lei, and Tat-Seng Chua. 2023b. Towards goal-oriented intelligent tutoring systems in online education. *CoRR*, abs/2312.10053.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Xuan Long Do, Bowei Zou, Shafiq R. Joty, Anh Tran Tai, Liangming Pan, Nancy F. Chen, and Ai Ti Aw. 2023. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10785–10803. Association for Computational Linguistics.

Chenhe Dong, Ying Shen, Shiyang Lin, Zhenzhou Lin, and Yang Deng. 2024. A unified framework for contextual and factoid question generation. *IEEE Trans. Knowl. Data Eng.*, 36(1):21–34.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1342–1352. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-unaware question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1919–1926. Association for Computational Linguistics.

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4853–4862. Association for Computational Linguistics.

Mark J Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6):1082–1116.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and QA over diverse inputs. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7879–7886. AAAI Press.

Michael Heilman and Noah A Smith. 2010. Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 11–20.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Myo-Kyoung Kim, Rajul A Patel, James A Uchizono, and Lynn Beck. 2012. Incorporation of bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American journal of pharmaceutical education*, 76(6).

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, BEA@NAACL-HLT 2015, June 4, 2015, Denver, Colorado, USA*, pages 154–161. The Association for Computer Linguistics.

Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving question generation with to the point context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3214–3224. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *CoRR*, abs/2309.12871.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6116–6124. Computer Vision Foundation / IEEE Computer Society.

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@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*, pages 284–290. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS 2022*.

Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9355–9367. Association for Computational Linguistics.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *CoRR*, abs/1905.08949.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2695–2709. Association for Computational Linguistics.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2096–2106. International Committee on Computational Linguistics.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 4339–4347. AAAI Press.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12047–12064. Association for Computational Linguistics.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023b. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *CoRR*, abs/2308.13259.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zichao Wang and Richard G. Baraniuk. 2023. Multiqg-ti: Towards question generation from multi-modal sources. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 682–691. Association for Computational Linguistics.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022. Towards human-like educational question generation with large language models. In *Artificial Intelligence in Education - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part I*, volume 13355 of *Lecture Notes in Computer Science*, pages 153–166. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: Fairytaleqa - an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 447–460. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2022. It is ai's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 731–744. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya 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.

Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang, and Lun-Wei Ku. 2022. Multi-vqg: Generating engaging questions for multiple images. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 277–290. Association for Computational Linguistics.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *CoRR*, abs/2311.09210.

Hongwei Zeng, Bifan Wei, Jun Liu, and Weiping Fu. 2023. Synthesize, prompt and transfer: Zero-shot conversational question generation with pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8989–9010. Association for Computational Linguistics.

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.

Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 218:1–218:21. ACM.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5073–5085. Association for Computational Linguistics.

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.

```
## 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: ...

## 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: ...

## 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: ...
```

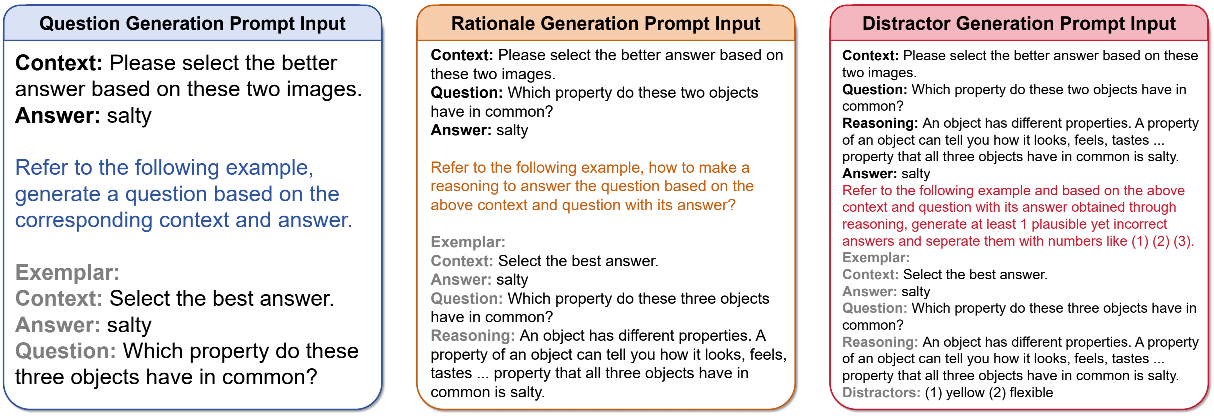


Figure 5: Schematic of three generation prompt constructions.

Method	Distinct-1	Distinct-2	Distinct-3	Distinct-4
ChatGPT 0 shot	12.69	41.43	62.86	75.35
ChatGPT 1 shot	12.90	41.16	62.86	75.54
ChatGPT 3 shot	10.94	36.28	57.16	70.64
VL-T5	7.23	18.59	28.96	37.24
MultiQG-TI	8.60	20.61	30.92	40.28
Multimodal-CoT	11.19	30.51	43.44	53.42
CoE	12.81	31.59	44.32	54.29

Table 7: Distinct-n scores of question generation.

Retrieval Strategy	Question Generation		Distractor Generation	
	B-4 ↑	R-L ↑	Acc ↓	R-L ↑
Random	70.4	79.3	78.6	58.3
Maximum	81.1	86.5	65.4	66.6
Summation	81.6	86.9	65.5	67.0

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

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.

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 influence 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:

$$\mathcal{I} = \underset{i\in\{1,2\ldots,N\}}{\arg\max}\ (Sim_t^i + Sim_a^i + Sim_q^i),$$

$$\begin{aligned} Sim_t^i &= \frac{(\mathcal{V}_t^i)^T\mathcal{V}_t}{\|\mathcal{V}_t^i\|_2\,\|\mathcal{V}_t\|_2}, \\ Sim_a^i &= \frac{(\mathcal{V}_a^i)^T\mathcal{V}_a}{\|\mathcal{V}_a^i\|_2\,\|\mathcal{V}_a\|_2}, \\ Sim_q^i &= \frac{(\mathcal{V}_q^i)^T\mathcal{V}_q}{\|\mathcal{V}_q^i\|_2\,\|\mathcal{V}_q\|_2}, \end{aligned} \tag{11}$$

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.

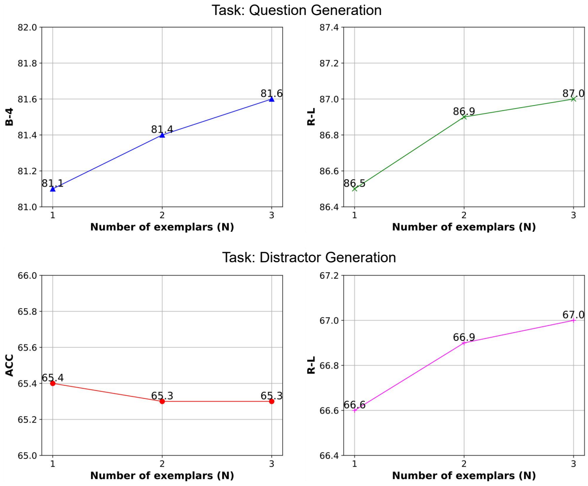


Figure 6: The overall performance with varying number of exemplars.

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.

Method	Question Generation		Distractor Generation	
	B-4 ↑	R-L ↑	Acc ↓	R-L ↑
Qwen-VL	**81.1**	**86.5**	**65.4**	**66.6**
LLaVA	80.7	86.0	70.1	60.3
InstructBLIP	80.5	86.0	69.7	60.0
mPLUG-Owl	80.3	85.8	72.3	58.0
VisualGLM-6B	51.2	73.2	77.5	39.6

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

E Guideline of Generation Quality Evaluation

We present the guideline of human evaluation for question and distractor generation quality in Figure 7.

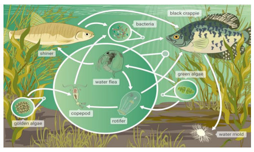
<table>
<tr><th colspan="2">Guideline of Generation Quality Evaluation</th></tr>
<tr><td colspan="2">This study aims to evaluate the quality of the question and distractor generation. Each case provides you with a context, image, answer and groundtruth. You need to evaluate the generated question and distractors from the following aspects.</td></tr>
<tr><th colspan="2">Case</th></tr>
<tr><td colspan="2">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. Image: 
Answer: copepod
Groundtruth Question: Which of the following organisms is the primary consumer in this food web?
Groundtruth Distractors: (1) black crappie (2) bacteria</td></tr>
<tr><th colspan="2">Question Evaluation</th></tr>
<tr><td colspan="2">➢ Readability: whether the generated questions are easy to read and understand for students of corresponding grade.</td></tr>
<tr><td>Options</td><td>1. Complete understanding 2. Quite understanding 3. Moderate understanding
4. Minor understanding 5. No understanding</td></tr>
<tr><td>Examples</td><td>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 of organism is the primary consumer in this food web is?” 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.</td></tr>
<tr><td colspan="2">➢ Appropriateness: whether the generated questions are semantically aligned with the corresponding subjects.</td></tr>
<tr><td>Options</td><td>1. Completely appropriate 2. Mostly appropriate 3. Fairly appropriate
4. Mostly inappropriate 5. Completely inappropriate</td></tr>
<tr><td>Examples</td><td>This case belongs to natural science.
1. “What is the largest fish in this food web?” shows fairly appropriateness in natural science domain.
2. “How many arrows are there in the picture?” is completely irrlevant to natural science.</td></tr>
<tr><td colspan="2">➢ Complexity: the level of reasoning or cognitive effort required to answer the generated question for students of corresponding grade.</td></tr>
<tr><td>Options</td><td>1. Pretty thought-provoking 2. Mostly thought-provoking 3. Fairly thought-provoking
4. Slightly thought-provoking 5. Completely unchallenging</td></tr>
<tr><td>Examples</td><td>1. “Which organism preys on golden algae in this food web?” is fairly thought-provoking and necessitates some reasoning effort to answer.
2. “What is the largest fish in this picture?” shows completely unchallenging to answer the question.</td></tr>
<tr><td colspan="2">➢ Engagement: whether the students find the questions engaging and have interest in answering the questions.</td></tr>
<tr><td>Options</td><td>1. Greatly engaging 2. Quite engaging 3. Fairly engaging 4. Slightly engaging 5. Not engaging</td></tr>
<tr><td>Examples</td><td>1. “Which organism preys on golden algae in this food web?” is fairly engaging since you need to answer this question by reasoning from both image and context.
2. “What does the food web model?” shows no engagement as the answer can be readily found in context without any significant cognitive effort.</td></tr>
<tr><th colspan="2">Distractor Evaluation</th></tr>
<tr><td colspan="2">➢ Consistency: whether the generated distractors are completely overlapping with the correct answer.</td></tr>
<tr><td>Options</td><td>1. Fully consistent 2. Mostly consistent 3. Moderately consistent
4. Mostly inconsistent 5. Fully inconsistent</td></tr>
<tr><td>Examples</td><td>1. “copepod” is completely overlapping with the correct answer.
2. “Copepoda” shows mostly consistentency with the correct answer.
3. “(1) black crappie (2) bacteria” are fully inconsistent with the correct answer.</td></tr>
<tr><td colspan="2">➢ Plausibility: whether the generated distractors are semantically relevant to the given context and question.</td></tr>
<tr><td>Options</td><td>1. Completely relevant 2. Mostly relevant 3. Moderately relevant
4. Mostly irrelavant 5. Completely irrelavant</td></tr>
<tr><td>Examples</td><td>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 irrelavant to the context.</td></tr>
<tr><td colspan="2">➢ Distinctiveness: the originality of the generated distractors in comparison to the groundtruth.</td></tr>
<tr><td>Options</td><td>1. Completely distinctive 2. Mostly distinctive 3. Moderately distinctive
4. A little distinctive 5. Fully overlapping</td></tr>
<tr><td>Examples</td><td>1. “(1) algae (2) bateria” 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.</td></tr>
</table>

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