Evaluating the Knowledge Dependency of Questions

Hyeongdon Moon[*†1] **Yoonseok Yang**[*†2] **Jamin Shin**[†]
Hangyeol Yu[1] **Seunghyun Lee**[1] **Myeongho Jeong**[1]
Juneyoung Park[1] **Minsam Kim**[†1] **Seungtaek Choi**[†1]
[1]Riiid AI Research
[2]UC Berkeley
hyeongdon.mun@riiid.co, yoonseok@berkeley.edu,
jayshin.nlp@gmail.com, {minsam.kim, seungtaek.choi}@riiid.co

Abstract

The automatic generation of Multiple Choice Questions (MCQ) has the potential to reduce the time educators spend on student assessment significantly. However, existing evaluation metrics for MCQ generation, such as BLEU, ROUGE, and METEOR, focus on the n-gram based similarity of the generated MCQ to the gold sample in the dataset and disregard their educational value. They fail to evaluate the MCQ's ability to assess the student's knowledge of the corresponding target fact. To tackle this issue, we propose a novel automatic evaluation metric, coined **Knowledge Dependent Answerability (KDA)**, which measures the MCQ's answerability given knowledge of the target fact. Specifically, we first show how to measure **KDA** based on student responses from a human survey. Then, we propose two automatic evaluation metrics, $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$, that approximate **KDA** by leveraging pre-trained language models to imitate students' problem-solving behavior. Through our human studies, we show that $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ have strong correlations with both (1) **KDA** and (2) usability in an actual classroom setting, labeled by experts. Furthermore, when combined with n-gram based similarity metrics, $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ are shown to have a strong predictive power for various expert-labeled MCQ quality measures. [1]

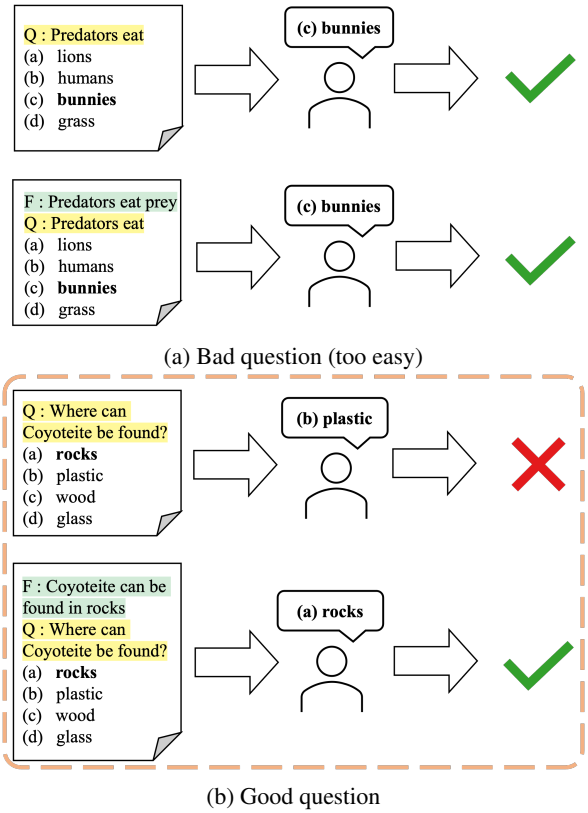

Figure 1: Problem formulation. A multiple-choice question designed to test the knowledge of a target fact must satisfy the following criterion: a question has to be answered by only a student who knows the fact, not a student who doesn't.

1 Introduction

Multiple-Choice Question (MCQ), comprised of a question stem, an answer, and a set of distractors, is one of the most widely used student assessment tools. Since manually generating exam-style questions is a complex, labor-intensive process that requires training, experience, and resources, automatic question generation (AQG) techniques were introduced (Kurdi et al., 2020). If automatic MCQ generation methods develop to a level that requires only minor adjustments by educators, it can meet the real-world demand for creating a massive amount of MCQ sets in seconds.

Despite the importance of AQG in educational purposes, the task did not receive much attention and was not actively introduced in the education field due to the limitation of evaluation methods. Previous AQG works mostly evaluate

[*] Equal Contribution.
[†] Corresponding authors. Work was done while all corresponding authors were at Riiid AI Research. Jamin Shin is currently at NAVER AI Lab.
[1] Our code is released here: https://github.com/riiid/question-score

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10512–10526
December 7-11, 2022

their methods based on how similar the generated questions/distractors are to the gold questions/distractors, using n-gram based similarity metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), or BERTScore (Zhang et al., 2019). However, all metrics mentioned above share the following limitations: (1) reliability of the evaluation depends on the quality of the reference dataset against which similarity is measured, and (2) n-gram-based similarity of a question does not directly measure the usability of the question as an assessment tool. Though a prior work (Nema and Khapra, 2018) proposed an alternative metric of answerability, the suggested metric does not account for whether the question's answerability crucially depends on knowledge of the target fact and also requires reference made by humans. AQG will be applied in the real-world much more easily if we address these limitations.

To this end, this paper suggests a new evaluation criterion for MCQ generation, coined **K**nowledge **D**ependent **A**nswerability (**KDA**). **KDA** measures whether a real student who got the problem wrong can choose the correct answer after gaining the knowledge the problem wants to test. **KDA** is based on the rationale that a well-designed MCQ should be answerable if the student knows the target fact being tested, which is briefly described in Figure 1.

However, as such human-dependent measurement is hard to be applied in real world scenario, we propose two automatic variants, $\mathbf{KDA_{cont}}$ and $\mathbf{KDA_{disc}}$. To automate the measurement of **KDA**, we regard the Pre-trained Language Models (PLMs) as question solvers that imitate students, based on the fact that large PLM's have strong question-answering abilities (Zhu et al., 2021; Roberts et al., 2020). To show the validity of the metrics, we ask two research questions: RQ1) When replacing students with PLMs, do the automated metrics ($\mathbf{KDA_{cont}}$ and $\mathbf{KDA_{disc}}$) behave similarly to **KDA** ? and RQ2) Can we use MCQs with high $\mathbf{KDA_{cont}}$ and $\mathbf{KDA_{disc}}$ in education field meaningfully?

To answer these research questions, we asked 116 students to solve 480 MCQs, which include both automatically generated questions by various methodologies and gold questions from three MCQ datasets. Among them, 96 questions were randomly selected, and a qualitative survey was conducted with experts to see if they were suitable for educational use. Through our experiments, we demonstrate that $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ both have a strong correlation not only with **KDA** as measured based on student responses but also with expert Likert score that measures the MCQ's overall usability in the classroom setting.

Our main contributions are as follows:

- We propose a novel, reference-free metric Knowledge-Dependent Answerability (**KDA**) that evaluates given MCQ's value as an assessment tool, and two automatic alternatives ($\mathbf{KDA_{disc}}$, $\mathbf{KDA_{cont}}$) of approximating **KDA** with PLMs.
- We validate the usability of $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ through extensive human studies.
- We release our code for **KDA**, in order to facilitate usage in public domain.

2 Related Work

2.1 Multiple-Choice Question Generation

Multiple choice question (MCQ) is comprised of a question, an answer, and a set of distractors. MCQ makes student assessment feasible, as it disambiguates a question by providing options to choose from (Rachmat and Arfiandhani, 2019). Especially, there have been multiple findings that active learning through question answering helps increasing learning gain of students (Crouch and Mazur, 2001; Koedinger et al., 1997; Wang et al., 2022). However, instructors suffer from generating high quality MCQs due to their limited resources.

Accordingly, there have been various attempts at automating MCQ generation. Here, we briefly review the automatic MCQ generation with regards to two important components: (1) question generation and (2) distractor generation.

Question Generation Question Generation (QG) is usually formulated as a task where an appropriate question must be generated given a reference document (Duan et al., 2017), where generated question can be answered from the reference document. For MCQ generation, target answer is usually given with the reference document, which serves as the answer for the generated question (Vachev et al., 2022). With the advent of deep learning, various neural networks have been used for QG: LSTM-based (Dong et al., 2018) and transformer-based (Laban et al., 2022; Hosking and Riedel, 2019).

Distractor Generation The goal of Distractor Generation (DG) is to receive a reference document, a question, and a target answer as inputs, and then output a set of distractors. There are mainly two classes of DG methods, namely, knowledge-based DG and language model-based DG.

Here, we regard knowledge-based DG as the superset of ontology-based DG and Knowledge-Driven DG. Historically, ontology-based DG was first proposed, which retrieves distractors that are similar to the answer according to the given domain-specific ontology (Alsubait, 2015). More recently, Knowledge-Driven DG has been proposed (Ren and Zhu (2021); KDDG), which uses a general-purpose knowledge base to select a pool of distractors, then rank the distractors using a feature-based model. In general, while knowledge-based DG provides plausible distractors semantically similar to the answer, they heavily rely on the quality of the knowledge base, and offer limited scalability.

To overcome the limitations of knowledge-based DG, recent works leveraged PLMs for DG (Chung et al., 2020; Vachev et al., 2022). Specifically, by fine-tuning a PLM to replicate gold distractors given the corresponding question and answer, the PLM learns how to make sensible distractors without relying on a knowledge base.

In this work, we employ T5 (Vachev et al., 2022) for both question generation and distractor generation due to its superior performance. Further, we implement KDDG and evaluate its performance using our proposed evaluation metric to represent knowledge-based DG.

2.2 Automatic Evaluation of MCQs

Following most works in language generation, prior works in QG and DG rely on n-gram similarity-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for automatic evaluation. BERTscore (Zhang et al., 2019) improves upon these metrics by performing n-gram matching with BERT-based contextualized embeddings. However, all of the aforementioned metrics only consider how similar the generated output is to the gold samples (Nema and Khapra, 2018), and fail to consider the value of generated questions as a student assessment tool.

To address this issue, (Nema and Khapra, 2018) considers Answerability of the question. In particular, they first collect human scores on whether an MCQ is answerable, then suggests four features for predicting the answerability score: relevant words, named entities, question words, and function words. Answerability metric is then defined as a learned weighted sum of the above features extracted from human, showing improved correlation with human judgement upon BLEU. Q-BLEU, which combines Answerability and BLEU, is shown to improve upon BLEU in terms of correlation with human judgment. However, this approach requires human annotations, and it is unlikely that results in one domain will extend to another. Also, the suggested answerability does not account for the source text given with the MCQ. (Wang et al., 2022) points out the low adoption of QG Systems in classrooms, requesting the QG system researchers to focus on educational needs.

3 Knowledge Dependent Answerability

To properly evaluate the question's answerability, we propose to measure **Knowledge Dependent Answerability (KDA)**. In particular, for each MCQ, we measure the proportion of students who answer the question correctly when they know the target fact being tested. Our rationale is that a good MCQ's answerability must crucially depend on the knowledge of the target fact. Then, in order to automatically measure KDA without human trials, we replace the students with Pre-trained Language Models based on their question-answering abilities (Zhu et al., 2021; Roberts et al., 2020).

3.1 Measuring KDA with student responses

Our goal is to measure the probability that a student will answer the MCQ correctly given that they know the target fact being tested. For this, we consider the participants who initially answer the target fact's pair question incorrectly as the ones who do not know the target fact. Then, we count the subset of those participants who answer the question correctly after the fact is shown. In other words, we measure the following:

$$KDA(q) = P(R^{q+f} = 1 \mid R^q = 0) \quad (1)$$
$$\approx \frac{\sum_j (1 - r_j^q) r_j^{q+f}}{\sum_j (1 - r_j^q)},$$

where R^q and R^{q+f} represent binary random variables for answer correctness before and after showing the fact, respectively. Here, lower-case variables represent sampled values, and j runs over all samples collected from human participants.

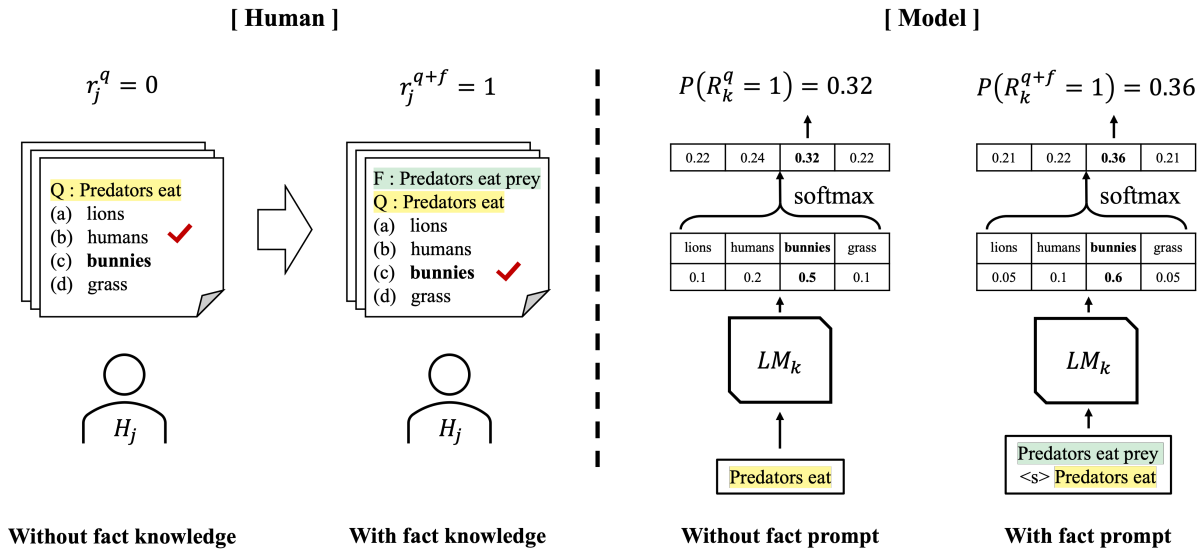

Figure 2: Flow diagram of human experiment and model inference. Left: human response phase and example. Each phase of human experiment tests 40 facts. After solving all questions without fact, each student solves all question again with corresponding fact. Right: inference of the language model. For several types of MCQ solver, we inference twice — with, and without the target fact.

3.2 Measuring KDA with Language Models

In order to approximate **KDA** (Equation 1), we use PLMs to emulate problem-solving students. In particular, in order to obtain r^q, we let the j-th language model to predict the answer given question q, as shown in Figure 2. We obtain r_j^{q+f} similarly, except we prompt the target fact in front of the question, as shown in Figure 2. We propose two versions of PLM-based **KDA**, namely, $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ (disc for *discrete*, and cont for *continuous*).

$$\mathbf{KDA_{disc}}(q) = \frac{\sum_j (1 - r_j^q) r_j^{q+f}}{\sum_j (1 - r_j^q)}. \quad (2)$$

$\mathbf{KDA_{disc}}$ exactly replicates Equation 1, but the binary values r_j^q and r_j^{q+f} are obtained from various PLM outputs instead of real students. Specifically, when the logit for the correct answer is the largest among the options, r_j^q (or r_j^{q+f}) equals 1, and otherwise 0.

$$\mathbf{KDA_{cont}}(q) = \frac{\sum_j P(R_j^q = 0) P(R_j^{q+f} = 1)}{\sum_j P(R_j^q = 0)}, \quad (3)$$

With language models, we can utilize probability outputs which may contain richer information compared to discretized values. Thus, we further propose the continuous version by replacing the binary values of Equation 1 with probability outputs from the language models. $\mathbf{KDA_{cont}}$ is interpreted as a weighted average of correctness probability $P(R_j^{q+f} = 1)$ across models, weighted by each model's probability of incorrect response without being shown the target fact. $\mathbf{KDA_{cont}}$ is also more stable than $\mathbf{KDA_{disc}}$, as the denominator of the Equation 2 can be zero if all model answer correctly to the question without fact.

4 Experiments

In this section, we first discuss the settings of our experiments to verify the efficacy of **KDA**, and then describe student and expert studies we conducted in detail. Lastly, we show our results for respective experiments.

4.1 Preliminaries

We designed our experiments to answer the following research questions:

- **RQ1**: Whether we can automatically measure **KDA** with $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ (Section 4.2)
- **RQ2**: Whether **KDA**, $\mathbf{KDA_{disc}}$, and $\mathbf{KDA_{cont}}$ correlate well with the judgments of real-world educators (Section 4.3)

To answer these questions, we need MCQ Generator Models to create questions on several datasets

	OBQA	TabMCQ	SciQ
Fact	predators eat prey	Urban sprawl creates thermal pollution	Plant hormones are chemical signals that control different processes in plants.
Question	Predators eat	What type of pollution does Urban sprawl create?	What chemical signals in plants control different processes?
Answer	bunnies	thermal pollution	plant hormones
Distractors	lions humans grass	air pollution radioactive pollution noise pollution	produce hormones nitrogen hormones Human Hormones

Table 1: Example questions from each dataset

and MCQ Solver Language Models to measure $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$.

MCQ Generator Models We divide MCQ Generation as only generating the distractors, and generating both distrators and question stem. 3 types of MCQ Generator models were used: two of them are of Distractor-only generation models and one generates both the question stem and distractors.

We first discuss the Distractor-only Generation models, which use fixed question stems provided from the datasets. For knowledge-based DG, we implemented KDDG (Ren and Zhu, 2021), which picks distractor candidates from the knowledge-base according to the topic distribution modeled by LDA (Blei et al., 2003) and then ranks them using learned features such as similarity to the answer. Following the authors' implementation, we selected Probase (Wu et al., 2012) as the knowledge base, and pair-wise LambdaMART ranker (Burges et al., 2011) as the ranker. As for PLM-based DG, we fine-tuned the `T5-Large model` (Raffel et al., 2020) on the RACE dataset (Lai et al., 2017), a large-scale reading comprehension MCQ dataset. To be specific, following (Vachev et al., 2022), we provided the question, answer, and the reference document as an input, then obtained three distractors. As shown in Table 3, the first distractors scored 46.59 BLEU1 on test data. We refer to this model as **T5DG**.

For *Question Generation (QG)*, we fine-tuned `T5-Large` (Raffel et al., 2020) models on each datasets to generate the question stems. The distractors are then generated with the above **T5DG**. This model, which we refer to as **QDG**, scored BLEU1 of 19.4, 53.3, 65.9 in OBQA, TabMCQ, and SciQ, respectively.

We prepared and evaluated MCQs generated by four different MCQ generation pipelines, as shown in Table 2. First, we evaluated MCQs generated by KDDG and T5DG based on human-created questions. Also, we evaluated MCQs whose question stem and distractors are both model-generated. Finally, we evaluated MCQs fully generated by humans, i.e., both questions and distractors are human-generated. More training details and results are included in the Appendix.

Generator	Distractors	Question Stem
Human	Human	Human
KDDG	KDDG	Human
T5DG	T5-DG	Human
QDG	T5-DG	T5-QG

Table 2: Baseline Models for QG/DG

	BLEU1	BLEU2	BLEU3	BLEU4
D1	46.59	38.33	34.31	32.02
D2	25.8	20.08	17.58	16.15
D3	28.33	23.07	20.73	19.46

Table 3: BLEU scores for T5-DG in RACE

MCQ Solver Language Models To calculate $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$, we prepare pre-trained language models as general-purpose MCQ solvers. In particular, in order to emulate students with various knowledge states and reasoning capabilities, we prepared 18 different PLM's of various types and sizes. For T5 models, we used the ones fine-tuned for Closed Book QA (Roberts et al., 2020). For other models, we fine-tuned the models with RACE dataset. A complete list of solver models can be found in Appendix.

	OBQA	TabMCQ	SciQ	All
$\mathbf{KDA_{cont}}$	**0.73****	0.16	0.17	0.74**
$\mathbf{KDA_{disc}}$	0.71**	**0.3****	0.05	**0.8****
BLEU	0.29**	0.14	0.16	0.26**
ROUGE-L	0.29**	0.14	**0.18***	0.27**
METEOR	0.28**	0.12	0.14	0.21**

Table 4: Pearson Correlation between $\mathbf{KDA}$ and other automatic evaluation metrics. Single (double) asterisk denotes p-value under 0.05 (0.01). Best correlation results are marked bold per dataset.

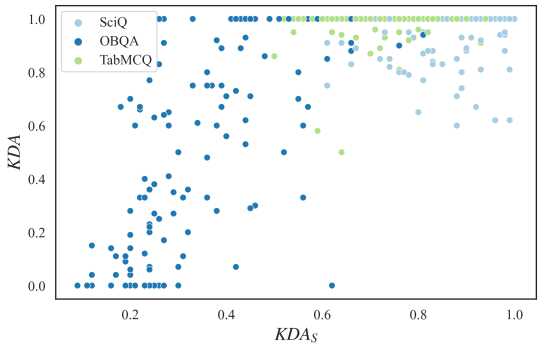

Figure 3: Scatter plot of the result. x-axis : $\mathbf{KDA_{cont}}$, soften $\mathbf{KDA}$ measured by language models. y-axis: $\mathbf{KDA}$, gold knowledge dependency measured by human solvers. The overall correlation between two metrics is 0.74

Datasets We used three real-world MCQ datasets for our experiments. Every MCQ contained in each dataset has three distractors and one answer.

- OpenBookQA (OBQA) (Mihaylov et al., 2018) is comprised of 5,957 elementary-level science questions. Answering the questions requires multi-step reasoning and commonsense knowledge.
- TabMCQ (Jauhar et al., 2016) contains 9,091 crowd-sourced MCQ science questions based on fact-based relation tables.
- SciQ (Welbl et al., 2017) contains 13,679 science exam questions about physics, chemistry, and biology.

4.2 RQ1: Correlation of KDA and $\mathbf{KDA}_*$

Setup To show that automatic evaluation metrics $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ indeed have strong correlations with $\mathbf{KDA}$, we conducted a large-scale human study with 116 participants. We randomly sampled 40 facts from each dataset (120 in total). Since we have four different MCQ generation pipelines for each fact, there were 4 different

Size and number of LMs	$\mathbf{KDA}$	Likert
LMs < 1GB (4 LMs)	0.65	0.36
LMs < 1.5GB (4 LMs)	0.71	0.38
LMs < 1.5GB (11 LMs)	0.73	0.39
All LMs (18 LMs)	0.74	0.43

Table 5: Pearson Correlation of $\mathbf{KDA_{cont}}$ with $\mathbf{KDA}$ and expert Likert score by the model size. We measured $\mathbf{KDA_{cont}}$ by using different models as solvers to show performance change along the number of language models used to calculate $\mathbf{KDA_{cont}}$. Each row of the table is obtained by using four small-size models only, the same numbers of larger-size models, 11 models under larger sizes, and the entire models.

questions asking for the same fact. Thus, we randomly mixed the questions from different models and splitted participants into four groups. More details can be found in the Appendix. The survey proceeded in the following stages, where participants had to answer 120 questions for each stage:

1) Question solving without fact: We first asked participants to solve the questions without showing them the relevant target facts.

2) Question solving with fact: Then, we asked participants to solve the questions with the relevant target facts given.

Results As seen in Table 4, both $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ show significantly higher correlation with $\mathbf{KDA}$ compared to n-gram based similarity metrics such as BLEU, ROUGE, and METEOR (0.74 and 0.80 respectively, with $p < 0.01$). This shows that ngram-based similarity metrics do not faithfully represent the MCQ's assessment value, while $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ can provide insight into the MCQ's assessment capabilities by considering language models' problem solving behavior.

Note that correlations may vary significantly across datasets due to different dataset characteristics. For example, although $\mathbf{KDA_{disc}}$ and $\mathbf{KDA_{cont}}$ correlate strongly with $\mathbf{KDA}$ for OBQA, the correlation is not as strong for TabMCQ and SciQ. This is because questions in OBQA require multi-step reasoning and commonsense knowledge (average correctness of 0.71 after human participants are shown the facts), while TabMCQ and SciQ mostly contain easy questions that are simple paraphrases of the relevant facts (corresponding human average correctness of 0.99 and 0.96, respectively). Therefore, while

Score	Descriptions
4	**[Strongly Agree]** This question can be readily used in the classroom.
3	**[Agree]** Despite some minor flaws, I'm willing to use this question to test the fact in a classroom
2	**[Disagree]** This question has some major flaws that needs to be revised for educational use in a classroom.
1	**[Strongly Disagree]** This question should be changed completely to be used in a classroom.

Table 6: Response types for the following question: On a scale of 1-4, how would you evaluate this question to be used in a classroom to test the fact below?

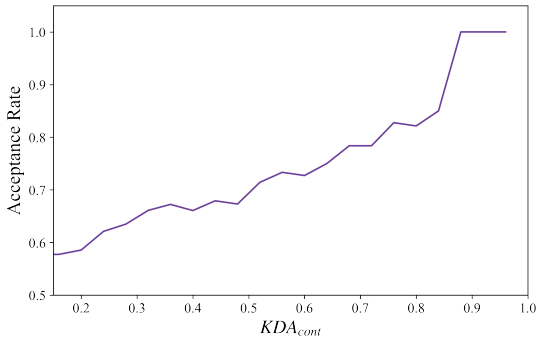

Figure 4: Cumulative graph showing how the acceptance rate of questions above a specific $\mathbf{KDA_{cont}}$ value changes. Given $\mathbf{KDA_{cont}}$ = 0.2, y-axis value is the ratio of question that got accepted among questions with $\mathbf{KDA_{cont}}$ over 0.2. 61%, 51%, 39%, and 19% of total questions has $\mathbf{KDA_{cont}}$ over 0.6, 0.7, 0.8, and 0.9, respectively. Thus, there are enough samples for each bin of $\mathbf{KDA_{cont}}$ threshold.

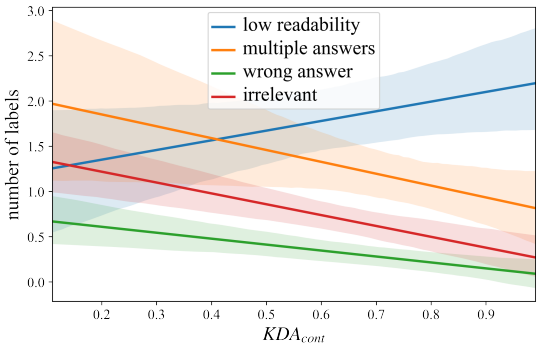

Figure 5: Regression graph of the number of expert labels by increasing $\mathbf{KDA_{cont}}$ of each question. The x-axis is $\mathbf{KDA_{cont}}$, and the y-axis is the number of flaw labels responded to by experts for each question with specific $\mathbf{KDA_{cont}}$ value. The light area corresponds to 95% CI.

language models may find MCQs from TabMCQ and SciQ relatively easy, human participants found the MCQs too easy, leading to a relatively weak correlation. This can be seen in Figure 3, where a lot of data points for OBQA and SciQ can be found around $y = 1.0$.

In addition, Table 5 shows the performance enhancement as models' size and the number of models increase. We believe a better correlation can be obtained when we use larger language models with improved reasoning capabilities to better imitate human problem solvers.

The results were obtained from a distribution in which a sufficient amount of samples existed even in a significant interval.

4.3 RQ2: Expert Judgment on $\mathbf{KDA_*}$

Setup To see whether **KDA** can measure the educational value of MCQs in a real-world classroom setting, we conducted a survey with previous or current secondary school science teachers. In particular, we randomly sampled 8 facts (32 questions) from each dataset (24 facts, 96 MCQs in total) among the facts that were selected for the human study. Given a pair of fact and an MCQ, we first asked the annotators to rate whether they would use the given MCQ to be used in a classroom to test the fact on a 4-point Likert scale (1: "Strong Reject", 2: "Weak Reject", 3: "Weak Accept", 4: "Strong Accept"), as shown in Table 6. Whenever an annotator rated an MCQ a score of 1 or 2, we further asked them why the given MCQ was unsatisfactory. They were able to choose one option between low readability, multiple answers, wrong answer, irrelevancy, and other. More experimental details can be found in the Appendix.

Results First, as shown in Figure 4, acceptance ratio increases as $\mathbf{KDA_{cont}}$ increases. Notably, we can see that **82%** of the questions which scored $\mathbf{KDA_{cont}}$ over **0.8** were accepted to be used in a classroom setting. This shows that $\mathbf{KDA_{cont}}$ can serve as a robust filtering metric to determine whether to use the generated question in a classroom setting or not.

We further examined whether $\mathbf{KDA_{cont}}$ can explain why an MCQ is unsatisfactory. Specifically, we fitted a linear regression model that takes

	$\mathbf{KDA}_*$	Others+$\mathbf{KDA}_*$	Others
Likert	0.33	**0.42**	0.21
Accept	0.41	**0.49**	0.19
Irrelevancy	0.22	**0.23**	-0.03
Low Readability	**-0.17**	-0.19	**-0.17**
Multi-Ans.	0.21	**0.35**	0.11
Wrong Ans.	**0.22**	0.18	-0.18

Table 7: Test set Pearson Correlation from Random Forest classifier. Others represent n-gram based similarity metrics (BLEU, ROUGE, METEOR) and $\mathbf{KDA}_*$ represent both $\mathbf{KDA}_{\mathbf{cont}}$ and $\mathbf{KDA}_{\mathbf{disc}}$. The results are averaged across 10 trials of 4-fold stratified cross-validation. Best result is marked in bold and second best result in underlined.

$\mathbf{KDA}_{\mathbf{cont}}$ as an input, and predicts the MCQ quality measures (Low Readability, Multiple Answers, Wrong Answer, Irrelevance). As seen in Figure 5, MCQs with higher $\mathbf{KDA}_{\mathbf{cont}}$ tend to be more relevant, while they are less likely to have multiple answers or wrong answer. However, $\mathbf{KDA}_{\mathbf{cont}}$ does not show prediction power on Low Readability. We believe that low readability is particularly hard to capture with $\mathbf{KDA}_{\mathbf{cont}}$: for example, students may find an MCQ hard to understand due to jargons and difficult vocabulary, while language models do not have such problem.

Finally, we examined which metrics best explain the given MCQ's quality annotated by experts—Likert scale and other binary labels of rejection such as Acceptance[2], Irrelevancy, Low Readability, Multiple Answers, and Wrong Answer. In particular, we trained a *Random Forest classifier*, which predicts MCQ's various quality measures based on the given input metrics. As shown in Table 7, using $\mathbf{KDA}_{\mathbf{disc}}$ and $\mathbf{KDA}_{\mathbf{cont}}$ outperforms using ngram-based similarity metrics (BLEU, ROUGE, METEOR) by a large margin for all but one quality measure of 'Readability'. Combining $\mathbf{KDA}$ metrics with similarity-based metrics, they show a strong synergy, reaching correlation of **0.49** with Accept labels.

4.4 Case Study

Here, we examine cases where $\mathbf{KDA}_{\mathbf{cont}}$ agrees with Expert Likert scores, aptly capturing the assessment quality of the question. As shown in Table 8, BLEU cannot take into account novel, yet convincing distractors (e.g., "wood" from #1) not present in the gold distractor set. However, $\mathbf{KDA}_{\mathbf{cont}}$ is able to take account the discriminative value of the distractors, and gives a high score to the MCQ, regardless of their resemblance to gold distractors. This is also demonstrated in #2, where QDG-generated question scored higher than the human generated question in Likert, $\mathbf{KDA}_{\mathbf{cont}}$, $\mathbf{KDA}_{\mathbf{disc}}$, but not in $BLEU$.

We also present the cases where $\mathbf{KDA}_{\mathbf{cont}}$ fails to measure the educational value of the given question.

Low KDA, High Likert: #3 shows the example of common types of question where $\mathbf{KDA}_{\mathbf{cont}}$ cannot be high. In order to get the question correctly, the student need to know that hexagon has six sides, which is not stated in the given fact. If the question requires some steps of reasoning, LMs would have relatively low correctness compared to the human solver, resulting in low $\mathbf{KDA}_{\mathbf{cont}}$ and $\mathbf{KDA}_{\mathbf{disc}}$. This can be resolved using LMs with better readig comprehension and reasoning abilities.

High KDA, Low Likert: #4 shows the case where $\mathbf{KDA}_{\mathbf{cont}}$ is high but Likert score is low. As shown in the table, the question asks the portion of the fact that is an assumption. While instructors can likely mark the question to have a low assessment value as it test a mere assumption, language model can not discern the quality of the fact the question is testing. Since SciQ is a crowd-source dataset, we noticed some questions don't assess meaningful target facts.

5 Conclusion

In this paper, we proposed Knowledge Dependent Answerability to measure the assessment value of the generated Multiple Choice Question. We formulated $\mathbf{KDA}$ as the probability that a student will solve the question correctly if the student know the target fact being tested. Then, we proposed $\mathbf{KDA}_{\mathbf{disc}}$ and $\mathbf{KDA}_{\mathbf{cont}}$ to approximate $\mathbf{KDA}$, treating PLMs as individual human solvers. Since both metrics are reference-free evaluation metrics, they can be applied without restrictions unlike previous metrics. They can be also applied as filters for question generation for educational use cases.

On 3 real-world MCQ datasets, $\mathbf{KDA}_{\mathbf{cont}}$ and $\mathbf{KDA}_{\mathbf{disc}}$ demonstrated a high correlation with $\mathbf{KDA}$ and the expert Likert score that measures the usability of the question as an assessment tool in a classroom setting. Notably, using $\mathbf{KDA}_{\mathbf{cont}}$ and

[2]Accept labels are given to questions with likert score higher than 2.5.

Examples	#	Model	Dataset	Question	Fact	Options	$\mathbf{KDA_{cont}}$	$\mathbf{KDA_{disc}}$	Likert	BLEU
KDA and Experts Agree	1	Human	TabMCQ	Where can Coyoteite be found?	Coyoteite can be found in rocks	rocks ✓ scissors spoons pens	0.90	1.00	3.29	100
	1	T5DG	TabMCQ	Where can Coyoteite be found?	Coyoteite can be found in rocks	rocks ✓ plastic wood glass	0.90	1.00	3.71	0
	2	Human	TabMCQ	A cave is formed by _.	A(n) cave is formed by weathering	weathering ✓ glacial erosion plate tectonics continental drift	0.77	0.87	3.00	100
	2	QDG	TabMCQ	A cave is formed by what process?	A(n) cave is formed by weathering	weathering ✓ glaciers volcanoes erosion	0.82	1.00	3.29	62.5†
KDA and Experts Disagree	3	KDDG	OBQA	Quartz crystals are made up of	a quartz is made of six-sided transparent crystals	hexagons ✓ oval square sphere	0.39	0.38	3.00	0
	4	KDDG	SciQ	Assume a molecule must cross a plasma membrane into what?	Assume a molecule must cross the plasma membrane into a cell... (excerpt)	cell ✓ electron tissue plasma	0.81	0.93	1.71	0

Table 8: Samples questions where **KDA** agrees or disagrees with Expert Likert scale. †: For QGD, BLEU for question stem was evaluated.

$\mathbf{KDA_{disc}}$ along with current n-gram based metrics drastically increased the overall correlation with these expert labels, as well as the prediction power of specific rejection reasons such as irrelevancy and wrong answer. We released a code and model weights to easily measure $\mathbf{KDA_{cont}}$ and $\mathbf{KDA_{disc}}$ for a given question and a fact pair. Future research may address expanding the metric's applicability to other types of assessment questions, such as short answer questions or multi-hop questions.

6 Limitations

In this section, we discuss limitations of our methods and experiments.

6.1 Prompt Can Bias Solvers' Decision

Our prompt-based method has limitations at estimating the student's knowledge on the fact, especially for multiple-answer situations. Suppose a question "Select an option that is in a liquid state at 20 Celcius" with multiple choices "water, orange juice, desk, and air", where the choice "water" is the only labeled answer, targeting a fact "Water is a liquid at 20 Celcius." However, as both "water" and "orange juice" are valid answers for the question, prompting the target fact can critically bias students' decision towards the only labeled answer "water," discounting the other valid answer "orange juice". In such cases, though the questions are of low quality, **KDA** cannot filter out the questions if the multiple answers are not labeled properly.

6.2 Too Easy Questions

Our experiments have an assumption that a well-designed MCQ is answerable if a student knows the target fact. However, there are edge cases where an MCQ question is answerable regardless of whether the student actually knew the fact. For instance, an answer can be easily found in the question stem itself or the distractors can be easily ruled out by simple topic irrelevance. In our experiment setting, this implies that there may have been students who are classified as knowledgeable, but didn't actually have the target knowledge as too easy questions cannot exactly distinguish the students' knowledge.

6.3 Difficulty of Measuring PLM's Ignorance

For PLM-based solvers, we do not have access to the huge training corpus in most cases, nor can we guarantee that the PLM *knows* a fact contained in the training corpus. We leave this as future work, as language models with language comprehension ability but without any knowledge are needed for such measure.

6.4 Low Agreement between Teachers

Table 9 shows the inter-rater agreement between two expert labelers as measured by Cohen's kappa. When we asked the teachers whether they would use the given MCQ in a real classroom setting (Section 4.3), the inter-rater agreement was relatively low, showing 0.2 on average among 7 annotators.

	obqa	tabMCQ	sciQ	human	qg+dg	dg	kddg	All
kappa	0.18	0.07	0.31	0.09	0.23	0.24	0.22	0.20

Table 9: Cohen's kappa coefficient for inter-annotator agreement. All 21 coefficient between 7 annotator were averaged.

We believe this is because the annotators had different subjective views on what makes a "good" MCQ. However, we noticed that reviewers have higher agreement on "bad" MCQs: questions with $\mathbf{KDA_{cont}}$ lower than 0.3 shows kappa coefficient over 0.3, and questions with $\mathbf{KDA_{cont}}$ lower than 0.2 shows kappa coefficient over 0.4.

6.5 Availability of PLMs for Low-Resource Languages

Effectiveness of $\mathbf{KDA_{cont}}$ and $\mathbf{KDA_{disc}}$ has only been tested in English questions with PLMs trained with English corpus. Since the metric depends on using PLMs that are trained to have a good reading comprehension ability, the use of the metric might be limited for low-resource languages.

7 Ethical Considerations

This study suggests that the question generation method can be applied in the real world, especially in the educational domain. Question for the purpose of assessment plays a vital role in the education system, and automatic question generation can reduce the cost of education providers. In this process, we expect our study to contribute to reducing inequality by increasing educational opportunities.

Despite such needs and efforts in AQG systems, automatically generated questions have shortcomings, even if they pass our proposed evaluation methodology. Since our method is designed to focus on knowledge dependency, several problems might remain in the filtered question, such as gender or racial bias (Hirota et al., 2022).

Our study aims to evaluate the absolute value of the question through language models rather than the existing psychometric-based relative evaluation method. Since there hasn't been much discussion about AQG systems in general, considerable attention and additional research is required before deploying AQG systems in real-world and evaluating them with our proposed metric.

Acknowledgements

We thank all of our human experiment and expert experiment participants for their time. We also thank the anonymous reviewers for their valuable feedback.

References

Tahani Mohammad Alsubait. 2015. *Ontology-based multiple-choice question generation*. Ph.D. thesis, The University of Manchester (United Kingdom).

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3615–3620.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Christopher Burges, Krysta Svore, Paul Bennett, Andrzej Pastusiak, and Qiang Wu. 2011. Learning to rank using an ensemble of lambda-gradient models. In *Proceedings of the learning to rank Challenge*, pages 25–35. PMLR.

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*, pages 4390–4400, Online. Association for Computational Linguistics.

Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American journal of physics*, 69(9):970–977.

Xiaozheng Dong, Yu Hong, Xin Chen, Weikang Li, Min Zhang, and Qiaoming Zhu. 2018. Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 213–223. Springer.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 866–874.

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Yusuke Hirota, Yuta Nakashima, and Noa García. 2022. Gender and racial bias in visual question answering datasets. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2278–2283.

Sujay Kumar Jauhar, Peter D. Turney, and Eduard H. Hovy. 2016. Tabmcq: A dataset of general knowledge tables and multiple-choice questions. *ArXiv*, abs/1602.03960.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kenneth R Koedinger, John R Anderson, William H Hadley, Mary A Mark, et al. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *NAACL-HLT*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959.

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*, pages 311–318.

Nadita Antania Rachmat and Puput Arfiandhani. 2019. "i use multiple-choice question in most assessment i prepared": Efl teachers'voice on summative assessment. *ETERNAL (English, Teaching, Learning, and Research Journal)*, 5(1):163–179.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2022. Leaf: Multiple-choice question generation. In *European Conference on Information Retrieval*, pages 321–328. Springer.

Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022. Towards process-oriented, modular, and versatile question generation that meets educational needs. In *NAACL*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 481–492.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Appendix

A.1 DataSet Preprocessing

OBQA, SciQ For OBQA and SciQ, we followed provided train-valid-test split. Questions containing "of above" or "of the above" are filtered out since they are exceptions to the problem formation. Train and valid splits are used to train the question generation model, and the questions we used in the human experiment and labeling are generated from the test set.

TabMCQ Since TabMCQ does not provide splits, and we divided it into 6:1:1, train, valid, test, respectively. Since the raw tabMCQ dataset does not provide evidence of the fact as a natural language form, the fact is generated by concatenating the expressions on the table. Regent tables 27 to 43 were omitted because they were not written in natural language formations. Facts were filtered if several instances were concatenated inside one column. Like other dataset settings, the question generation model is trained using train validation splits, and the questions from the test split are used for the evaluation.

A.2 Question Generator Training Details

QG, DG T5-large model was used, with max target token length of 512, target token length of 64, early stopping patience of 10 epochs, learning rate of 0.00001 with AdamW optimizer, and batch size of 24 (4 A100 machines in DDP, batch size of 6 for each machine).

KDDG implementation is done by description and open source of its original work. (Ren and Zhu, 2021) Since the source code and the resources are partially accessible and the appendix part, which is mentioned to contain the experiment settings, was missing, we re-implemented many parts of the code. Probase, which is renamed to Microsoft Concept Graph [3], is used as the knowledge base. Latent Dirichlet Allocation(LDA) for topic modeling uses Gensim library[4] and is trained on the processed Wikipedia corpus provided by the library. Feature Extractor for distractor selector were implemented except omitted *Contextual Embedding similarity* and *Web-search Score* at the code. LambdaMart [5] is used as a Ranker, while the number of trees is 2, and the learning rate is 0.1.

[3] https://concept.research.microsoft.com/Home/Download

[4] https://radimrehurek.com/gensim/

[5] https://github.com/lezzago/LambdaMart

A.3 Human Experiment Details

Among the generated questions, we filtered out incompletely generated questions that have less than 3 distractors. Among all the facts, we randomly sampled 40 facts from each dataset. As visualized in 2, students are asked to answer the question with and without the corresponding fact. Total 116 student joined the experiment questionnaire made by Typeform [6]. We split the experiment group into 8 sessions, and each group takes 2 hours to solve the questions following the instructions. Due to COVID-19, we conducted the experiment remotely, while a non-face-to-face video call solution Zoom [7], the experiment was conducted by checking the participant's face and the participant's screen. (fact validation)

In addition to the experiment setting described in the 4, we asked the participants whether they knew the relevant facts as the prior knowledge after answering the questions without facts and before answering with facts. We showed the fact independently and let the students choose 'yes' or 'no' to answer the question, "Did you previously know the below fact?" Its purpose was to resolve the limitation of the prompt-based knowledge-providing method and utilize the response as a gold label of the direct knowledge state of students. However, we can not find a meaningful explanation for the relationship between this label and our automatic metrics or human Likert score. Considering the average correctness of TabMCQ questions with fact is 0.99, we can expect that TabMCQ has very high relevance between the given fact and the question. Surprisingly, over 33% of students answered the question incorrectly but responded knowing the given fact and vice versa. Further research is needed to explain the difference between having knowledge and being presented with a prompt.

A.4 Expert Labeling Details

The experiment was conducted with 7 high school science teachers. During the 1 hour and 30 minute experiment, the teacher received a 150$ amazon gift card as a reward. A total of 8 types of facts and 32 questions were presented for each dataset, and a questionnaire was conducted on a total of 96 questions. For each multiple-choice question, the questionnaire was asked whether to use the Likert scale in the actual educational field, and the

[6] https://www.typeform.com/

[7] https://zoom.us/

question to choose the most prominent reason not to use the problem was answered.

A.5 MCQ Solver implementation

The entire model list are as follows.

- T5-cbqa-small, T5-cbqa-large, T5-cbqa-xxl (Roberts et al., 2020)
- bert-base, bert-large (Kenton and Toutanova, 2019)
- Roberta-base, Roberta-large (Liu et al., 2019)
- MPNet (Song et al., 2020)
- SciBERT (Beltagy et al., 2019)
- XLNet-base, XLNet-large (Yang et al., 2019)
- BioBERT-base, BioBERT-large (Lee et al., 2020)
- DistillBERT-base, DistillRoberta-base (Sanh et al., 2019)
- ALBERT-xl, ALBERT-xxl (Lan et al., 2019)
- MatsciBERT (Gupta et al., 2022)

Except for T5 models trained on closed-book question answering, all models are fine-tuned to the RACE dataset to answer MCQs. AdamW optimizer was used with a learning rate of 1e-5 for base models and 1e-6 for large models. Early stopping patience of 5 epochs was used. Base models were trained with a batch size of 32, and large models were trained with a batch size of 16. A cloud instance is used on google cloud service[8] to train the solver models. Overall, the cloud instance cost for training is about to 14k$, while the pricing policy is 'on demand,' which costs much higher than the preemptive instance.

For Table 5, used models are as follows.

- LMs < 1GB (4 LMs): DistillBert-base, DistillRoberta-base, Albert-xl, T5-cbqa-small
- LMs < 1.5GB (4 LMs): bert-base, Roberta-base, XLNet-base, MPNet
- LMs < 1.5GB (11 LMs): DistillBert-base, DistillRoberta-base, Albert-xl, T5-cbqa-small, bert-base, Roberta-base, BioBERT-base, SciBERT, MatsciBERT, XLNet-base, MPNet

[8] https://cloud.google.com/gcp

	OBQA	TabMCQ	SciQ	All
Human	-0.57	-0.18	0.06	-0.01
KDDG	0.01	0.13	0.75*	0.42*
DG	0.59	-0.01	0.65	0.61**
QDG	0.13	0.26	0.58	0.64**
DGen models	0.32	0.16	0.65**	0.51**

Table 10: Correlation of $\mathbf{KDA_{cont}}$ and Human Likert Score per generate model. DGen models are counting both KDDG and DG models to figure the performance on evaluating distractor generation models.

	avg. Rq	avg. Rf	avg. Rq+f
OBQA	0.58	0.80	0.71
TabMCQ	0.53	0.42	0.99
SciQ	0.56	0.42	0.96

Table 11: Average Correctness of Datasets. TabMCQ and SciQ are notably higher than OBQA.

Sub Metric	Model Count (Total Size)	**KDA** (Valid)	Likert (Test)
$\mathbf{KDA_{small}}$	4 (3.5GB)	0.740	0.377
$\mathbf{KDA_{large}}$	10 (19.2GB)	0.784	0.421

Table 12: Two sub metrics of $\mathbf{KDA_{cont}}$ developed for the convenience of use. $\mathbf{KDA_{small}}$ uses T5-cbqa-small, ALbert-xl, MPNet, SciBert to calculate $\mathbf{KDA_{cont}}$, and $\mathbf{KDA_{large}}$ uses T5-cbqa-small, T5-cbqa-large, ALbert-xl, MPNet, SciBert, bert-base, BioBert-base, Roberta-base, Roberta-large, XLNet-large.

A.6 Random Forest Implementation Detail

Among tree depths of 2 to 4, depth 2 was selected as Others (BLEU, ROUGE, METEOR) showed the best performance in that depth. The test set was composed of 2 key facts per dataset (8 questions per dataset, total 24 questions). The result was averaged across 10 trials of 4-fold stratified cross-validation. No hyper-parameters were tuned from the default sklearn random forest setup.

A.7 Sub Metrics

Since utilizing all models used to calculate $\mathbf{KDA_{cont}}$ or $\mathbf{KDA_{disc}}$ needs huge compute resources, we provide two variant metrics, $\mathbf{KDA_{small}}$ and $\mathbf{KDA_{large}}$. The subset of entire models is selected under the constraints of the number and total sizes of models. We used questions without an expert label as a validation set to pick a combination of the highest correlation to $\mathbf{KDA_{cont}}$and measure correlation to an expert Likert score as a test, and correlation is at Table 12.

Examples	Eg1	Eg2	Eg3	Eg4
Dataset	OBQA	OBQA	OBQA	OBQA
QG model	KDDG	KDDG	KDDG	DG
Fact	a beach ball contains gas	water is in the solid state, called ice , for temperatures between 0 and 0 F	friction acts to counter the motion of two objects when their surfaces are touching	friction acts to counter the motion of two objects when their surfaces are touching
Question	Which would you likely find inside a beach ball?	Global warming is lowering the world's amount of	When it's flying, a plane has no friction with the	When it's flying, a plane has no friction with the
answer	air	ice	ground	ground
options	food gas water	snow water air	power air water	air water sky
gold_options	steam water cheese	hurricanes carbon dioxide ocean levels	wings clouds air	wings clouds air
Likert	2.57	2.57	2.71	3.0
$\mathbf{KDA_{cont}}$	0.11	0.38	0.27	0.29
$\mathbf{KDA_{disc}}$	0.08	0.36	0.40	0.54
BLEU	33.3	0.00	33.3	33.3

Table 13: We report all cases of $\mathbf{KDA_{cont}} < 0.4$ and $\text{human}_{\text{likert}} > 2.5$

Examples	Eg1	Eg2	Eg3	Eg4
Dataset	TabMCQ	TabMCQ	TabMCQ	TabMCQ
QG model	QG+DG	QG+DG	DG	DG
Fact	water is an insulator of electricity	air is an insulator of electricity	warm is a term that can describe air temperature	water is an insulator of electricity
Question	Water is an insulator of what?	Air is an insulator of what?	What does the term warm describe?	Water is an insulator of what?
answer	electricity	electricity	air temperature	electricity
options	cold heat warmth	heat cold warmth	precipitation wind speed cloud cover	heat warmth cold
gold_options	wind air heat	heat water metal	wind speed air pressure optical phenomenon	wind air heat
Likert	2.14	2.42	2.29	2.00
$\mathbf{KDA_{cont}}$	0.84	0.82	0.84	0.84
$\mathbf{KDA_{disc}}$	0.93	0.93	0.86	0.93
BLEU	33.33	33.33	33.33	33.33

Table 14: We report all cases of $\mathbf{KDA_{cont}} > 0.8$ and $\text{human}_{\text{likert}} < 2.5$