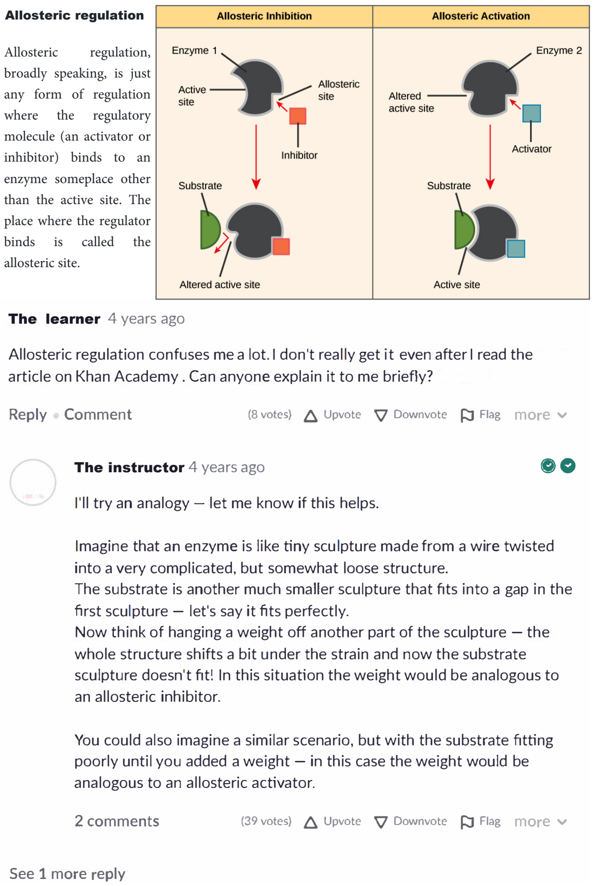
KhanQ: A Dataset for Generating Deep Questions in Education

Huanli Gong[1] **Liangming Pan**[2] **Hengchang Hu**[2]
[1]The Ohio State University, Columbus, OH, USA
[2]School of Computing, National University of Singapore, Singapore
gong.545@osu.edu
{liangmingpan, hengchanghu}@u.nus.edu

Abstract

Designing in-depth educational questions is a time-consuming and cognitively demanding task. Therefore, it is intriguing to study how to build Question Generation (QG) models to automate the question creation process. However, existing QG datasets are not suitable for educational question generation because the questions are not real questions asked by humans during learning and can be solved by simply searching for information.

To bridge this gap, we present KhanQ, a challenging dataset for educational question generation, containing 1,034 high-quality learner-generated questions seeking an in-depth understanding of the taught online courses in Khan Academy. Each data sample is carefully paraphrased and annotated as a triple of 1) *Context*: an independent paragraph on which the question is based; 2) *Prompt*: a text prompt for the question (*e.g.*, the learner's background knowledge); 3) *Question*: a deep question based on *Context* and coherent with *Prompt*. By conducting a human evaluation on the aspects of appropriateness, coverage, coherence, and complexity, we show that state-of-the-art QG models which perform well on shallow question generation datasets have difficulty in generating useful educational questions. This makes KhanQ a challenging testbed for educational question generation for further investigation.



Allosteric regulation

Allosteric regulation, broadly speaking, is just any form of regulation where the regulatory molecule (an activator or inhibitor) binds to an enzyme someplace other than the active site. The place where the regulator binds is called the allosteric site.

The learner 4 years ago

Allosteric regulation confuses me a lot. I don't really get it even after I read the article on Khan Academy . Can anyone explain it to me briefly?

The instructor 4 years ago

I'll try an analogy – let me know if this helps.

Imagine that an enzyme is like tiny sculpture made from a wire twisted into a very complicated, but somewhat loose structure.
The substrate is another much smaller sculpture that fits into a gap in the first sculpture – let's say it fits perfectly.
Now think of hanging a weight off another part of the sculpture – the whole structure shifts a bit under the strain and now the substrate sculpture doesn't fit! In this situation the weight would be analogous to an allosteric inhibitor.

You could also imagine a similar scenario, but with the substrate fitting poorly until you added a weight – in this case the weight would be analogous to an allosteric activator.

Figure 1: A real human-raised question after learning the "allosteric regulation and feedback loops" course in Khan Academy.

1 Introduction

Question asking has long been considered a fundamental cognitive process in learning. An ideal learner should be an active, self-motivated, creative, inquisitive person who asks deep questions and searches for answers to thought-provoking questions, usually in the form of *Why*, *Why-not*, *How*, *What-if*, etc (Otero and Graesser, 2001). For example, Figure 1 shows a question raised by a learner after learning "allosteric regulation and feedback loops" in Khan Academy[1], a well-known online education platform. Given that question-asking is a hallmark of human intelligence, it is intriguing to study whether we can endow machines with the ability to ask deep and to-the-point questions.

While there have been considerable advances made in the field of Question Generation (QG), the current research is still far from achieving human-like question-asking. One major obstacle is the lack of a suitable and clean dataset that can represent

[1] https://www.khanacademy.org/

Proceedings of the 29th International Conference on Computational Linguistics, pages 5925–5938
October 12–17, 2022.

real-life human-raised questions, such as the question shown in Figure 1. Existing QG works (Duan et al., 2017; Zhao et al., 2018; Pan et al., 2020; Back et al., 2021) typically focus on generating factoid questions relevant to one fact obtainable from a single sentence. They are shallow because they do not reflect the creative *human cognitive process* in question-asking such as inferences, multi-step reasoning, synthesis, critical evaluation, and generalization.

To bridge this gap, we desire a new dataset with questions that satisfy two conditions: 1) they are real questions asked by humans during learning, and 2) they are deep questions that require high-level cognitive skills to raise. We choose to collect data from Khan Academy, where the questions are raised by real learners after watching course videos or reading course materials, as shown in Figure 1. These questions represent what learners are naturally interested in, and they are rarely shallow questions confined to a narrow scope of context.

We collect the question-answer pairs from Khan Academy and rewrite them as a triple of 1) *Context*: an independent paragraph on which the question is based; 2) *Prompt*: a text prompt for the question (*e.g.*, the learner's background knowledge); 3) *Question*: a deep question based on *Context* and coherent with *Prompt*. Following the common setting of question generation, questions are based on the information in the context. However, many valid questions can be asked from the given context. To facilitate the evaluation and to guide the generation, we also provide *Prompt* which describes the questioner's background knowledge (*e.g.*, "In my understanding, ...") or certain conditions for the question (*e.g.*, "When ... happens, ..."). Given both *Context* and *Prompt* as inputs, we test the model's ability to generate a consistent question with both the context and the prompt. We design a rigorous data annotation procedure to ensure that each sample in KHANQ satisfies the following conditions: 1) *Question* involves deep reasoning instead of simply searching for information; 2) *Context* contains enough information to derive *Question*; 3) *Question* is coherent with *Prompt*. After careful annotation, we collect 1,034 (*Context*, *Prompt*, *Question*) triples to form the KHANQ. An example is given in Figure 2.

We further evaluate the depth of the questions in KHANQ. We find that KHANQ is dominated by deep questions that are represented in the form of *Why*, and *How*. In contrast, SQuAD 2.0 (Rajpurkar et al., 2018) and HotpotQA (Yang et al., 2018) are dominated by shallow questions represented in the form of *what*. We further classify these questions based on their reasoning type following the criteria in (Cao and Wang, 2021). The results show that 51.58% of questions in KHANQ involve deep reasoning.

Context: The molecules in the water have a range of kinetic energies, from low to high. Some of the molecules have sufficient kinetic energy to break out and to enter the air. In coastal areas there are also factors. For example, waves churn up the water and give rise to very fine droplets that can get carried in the wind.
Prompt: Water have strong cohesion
Question: How does water evaporate?

Figure 2: A data sample in our KHANQ dataset.

We evaluate the performance of five text generation models, which have achieved state-of-the-art question generation performance on the SQuAD dataset: BertGeneration (Rothe et al., 2020), GPT-2 (Radford et al., 2019), BART (Lewis et al., 2020), Google-T5 (Raffel et al., 2020), UniLM (Dong et al., 2019). By conducting both automatic and human evaluation, we find that although the abilities of BART and Google-T5 to reason in settings and ask deep questions are better than other models, the question quality is still much lower than the human level. The results show that the KHANQ is a challenging testbed for generating human-like deep questions in education.

2 Related Work

Question Generation (QG) aims to automatically generate questions from textual input. Earlier work relied on syntactic transformations to convert declarative sentences into questions (Chali and Hasan, 2015; Heilman and Smith, 2010). Recent neural models (Zhou et al., 2019; Krishna and Iyyer, 2019; Sun et al., 2018) rely on sequence-to-sequence models to generate questions from a given sentence or paragraph by considering the focus, type, and general specific relations of the question. However, these approaches only involves generating shallow factoid questions, which are typically trained and evaluated on the SQuAD (Rajpurkar et al., 2016) dataset. SQuAD is insufficient to evaluate deep QG because more than 80% of its questions are only relevant to information confined to a single sentence (Du et al., 2017).

With the advent of pretraining language models, the challenge of generating single-hop questions similar to SQuAD have largely been addressed. QG research has started to generate more complex questions that require multi-hop reasoning (Tuan et al., 2020; Pan et al., 2020; Xie et al., 2020; Yu et al., 2020), benchmarking on the HotpotQA (Yang et al., 2018) dataset. However, for questions in HotpotQA, the reasoning is often evident from the surface form of the question, simplifying the QG task. For example, Pan et al. (2021) show that HotpotQA-style multi-hop questions can be generated by composing single-hop questions through templates. Different from SQuAD and HotpotQA, questions in our KHANQ dataset are asked by real course learners, thus having a wider variety in both question forms and reasoning types.

Recently, a few works started to work on generating real human-raised questions. For example, Cao and Wang (2021) collect real questions from online forums (Reddit and Yahoo). Gupta et al. (2019) collect questions posted by customers on Amazon product pages. However, questions collected in above works are noisy, with few in-depth questions, since they fail to carefully filter and validate the questions. Compared to them, questions in KHANQ are carefully filtered and annotated, providing a more clean testbed for deep question generation. Among these works, LearningQ (Chen et al., 2018) is closest to us, which also collects questions from Khan Academy. However, our work has three key differences. First, LearningQ is more focused on the educational nature of the questions. They filter questions based on whether they are useful for learning, whereas we focus more on the depth of the question, keeping only the questions that involve deep reasoning. Second, we use prompts to give the models additional guidance on what information to focus on when generating. Third, the contexts used by LearningQ are entire articles or videos, causing most of the sentences in the contexts being irrelevant to the target question. In contrast, the contexts we use are answers that contain comprehensive explanations of the knowledge points relevant to the question.

3 Data Collection and Annotation

3.1 Text Sources

Khan Academy is an online education institution that provides free teaching materials. Online course learners are encouraged to ask questions in the corresponding forum to clarify their confusions after they learned each section of the course, as shown in the example in Figure 1. Therefore, these questions usually reflect a high-level comprehension and cognition of the course contents, which makes them a suitable data source for building a dataset for deep question generation. We chose to collect data from the courses in the science domain because question-asking is more active in the science-related courses than others. Learners asked more questions and most of them have been answered by tutors.

As shown in Table 1, we collected a total number of 1,284 course sections from 11 different areas under the science domain. Each course section consists of the course material (an article) written by the tutor and a discussion forum. We filtered out those course sections that have no question in their discussion forums. In total, we collected 100,908 question-answer pairs.

area	number
High school biology	209
AP/College Biology	296
High school chemistry beta	4
AP/College Chemistry	307
Organic chemistry	290
High school physics	82
AP/College Physics 1	4
AP/College Physics 2	19
AP/College Environmental science	18
Cosmology and astronomy	2
Electrical engineering	53

Table 1: The number of collected course sections for each area in the science domain.

3.2 Data Annotation

To build a clean and high-quality dataset for deep question generation, we then designed a rigorous data annotation procedure to filter out noisy data in our collected question-answer pairs. After filtration and annotation, each data sample is annotated as a triple of: 1) *Context*: an independent paragraph which the question is based on; 2) *Prompt*: a text prompt for the question (*e.g.*, the learner's background knowledge); 3) *Question*: a deep question based on *Context* and *Prompt*. Compared with the original noisy question-answer pairs collected from forums, data annotation provides a standardized and clean benchmark to train and evaluate the question generation models.

Figure 3 summarizes the major steps of our data annotation. First, the *Context* of the question should cover the knowledge points that the question are based on. To provide the context, we ask annotators to revise the answer provided by the tutors of the course. We find that most answers contain comprehensive explanations of the knowledge points relevant to the question from the course. Therefore, they are suitable to serve as the context for the question. We asked the annotators to remove the answer-tone phrases (*e.g.*, "yes", "this is because") and conversational language (*e.g.*, "good question", "hope it helps") to make the answer context-independent and self-contained.

For the original questions raised by learners, they often contain conditional clauses, prepositional phrases, and other conditions to limit the scope of the question or to provide some background information that expresses the learner's own understanding or opinions. As shown in Figure 3, we ask human annotators to separate out this information from the original question to serve as *Prompt* of the question. The prompt provides additional guidance for the question generation model in knowing which information to focus on when generating; otherwise, the QG model tend to generate questions without specific target (e.g., "Can you explain this part again?"). The *Question* then comes from the remaining part of the original question that seeks for new information based on the prompt. Appendix A gives detailed data filtration criteria, specific data annotation guidelines for each step, and examples of annotation.

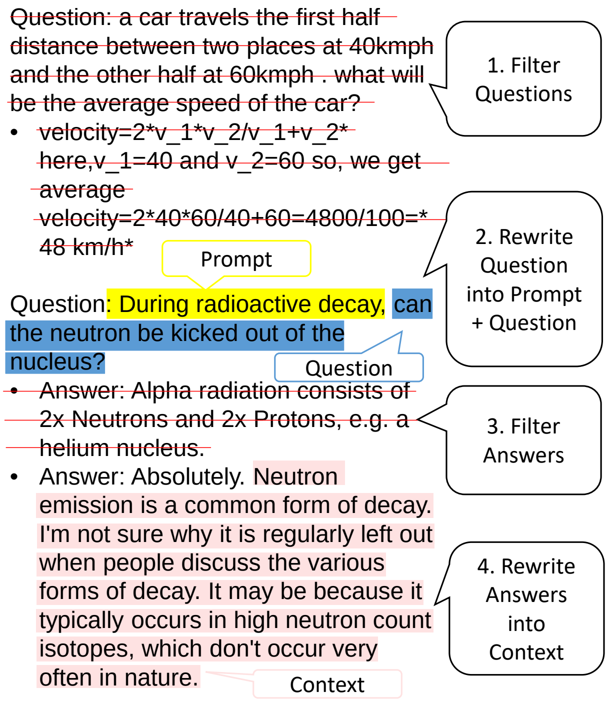

~~Question: a car travels the first half distance between two places at 40kmph and the other half at 60kmph . what will be the average speed of the car?~~
~~velocity=2*v_1*v_2/v_1+v_2* here,v_1=40 and v_2=60 so, we get average velocity=2*40*60/40+60=4800/100=* 48 km/h*~~
1. Filter Questions
Prompt
Question: During radioactive decay, can the neutron be kicked out of the nucleus?
Question
2. Rewrite Question into Prompt + Question
~~Answer: Alpha radiation consists of 2x Neutrons and 2x Protons, e.g. a helium nucleus.~~
3. Filter Answers
Answer: Absolutely. Neutron emission is a common form of decay. I'm not sure why it is regularly left out when people discuss the various forms of decay. It may be because it typically occurs in high neutron count isotopes, which don't occur very often in nature.
4. Rewrite Answers into Context
Context

Figure 3: Data annotation pipeline, consisting of four steps: 1) Filter Questions; 2) Rewrite Question into Prompt + Question, 3) Filter Answers, and 4) Rewrite Answers into Contexts.

3.3 Quality Control

To control the quality of the annotation, we require that annotators have a US high school diploma or equivalent to demonstrate sufficient background knowledge to understand the questions. To ensure that annotators understand the annotation procedure, we check those works and give feedback when an annotator has completed the first 10% of the tasks. The annotator needs to redo those annotations based on the feedback. This process is repeated until all 10% of the tasks are approved. During the annotation process, we also spot-check the work submitted by the annotators. If the pass rate does not reach 85%, the annotator needs to be retrained to prevent them from deviating from the task definition. Based on the above criteria, we hired a total of six annotators. The whole annotation process took two months. 30,000 original question-answer pairs are examined and 1,217 of them are annotated as valid samples.

After the annotation, we hired two workers to validate the quality of annotation. We find that: in 140 data samples, the context cannot provide the information needed to drive the question. In 19 data samples, the prompt cannot constrain the direction of the question. In 25 data samples, The question is incomplete, incomprehensible, not even a question at all. 85% of the annotated data samples meet all the requirements, which gives us 1,034 samples to form our final KHANQ dataset.

4 Data Analysis

4.1 Dataset Statistics

The final dataset consists of 1,034 high-quality data samples, in which 515 samples come from the field of Biology, 401 from the field of chemistry, 88 from the field of physics, 19 from the field of electrical-engineering, 7 from the field of environmental-science, and 4 from the field of cosmology-and-astronomy. The average length of *Context*, *Prompt*, and *Question* are 84.62 words, 14.12 words, and 10.74 words, respectively.

4.2 Question types by words

To investigate the depth and diversity of questions in KHANQ, we classify questions based on the first two words in the question and compare them to other commonly-used question generation datasets: SQuAD 2.0 (Rajpurkar et al., 2018) and HotpotQA (Yang et al., 2018), as shown in Table 2.

KHANQ	%	SQuAD	%	HotpotQA	%
Why does	3.8	what is	8.5	What is	5.0
How does	3.6	what was	5.3	Who was	2.1
Why is	3.4	how many	4.9	What was	2.0
Does the	3.0	when did	3.1	In what	1.8
Why do	2.8	in what	2.9	When was	1.7
What is	2.8	what did	2.8	Who is	1.6
How do	2.3	when was	2.1	How many	1.0
How to	2.2	who was	2.1	In which	0.9
How is	2.2	what does	1.7	What year	0.9
Is it	2.2	what are	1.7	Are both	0.9

Table 2: Most frequent leading bigrams in SQuAD 2.0, HotpotQA and KHANQ

According to (Craig et al., 2000), questions that are represented in the form of *Why*, and *How* are likely to be deep questions. Table 2 shows that KHANQ is dominated by these questions. In contrast, SQuAD 2.0, and HotpotQA are dominated by shallow questions represented in the form of *what*.

4.3 Question types by reasoning

To gain a better insight of the characteristics of the questions, we manually analyzed a sample of 95 different questions from KHANQ. We classify these questions following the criteria in (Cao and Wang, 2021). We summarize the most common types of questions in KHANQ and their corresponding examples as follows.

PROCEDURAL questions In 21.05% of the questions we inspected, the questioners asked for the procedures, tools, or methods by which a particular outcome is achieved. Most of these questions begin with *How*, followed by an auxiliary verb, a modal verb, or *to*, which indicates that the questioner is reasoning about the steps of action.

• *How to determine the oxidation state of oxygen?*

• *How does the body know what to send down the esophagus and what to send down the trachea?*

CAUSE questions In 18.95% of the questions we inspected, the questioners are asking for the cause or reason for an event or a concept. Most of these questions begin with *Why*, followed by an auxiliary verb, a modal verb, or their negative form, which indicates that the questioner is seeking the reason behind a phenomenon.

• *Why are the electrons mapped out in such an orderly way?*

• *Why don't the oxygen atoms in the air bond to the water molecules on the surface and pull on water molecules?*

VERIFICATION questions In 15.79% of the questions we inspected, the questioners asked for the truthfulness of an event or a concept. Most of these questions are general questions that begin with be verbs, auxiliary verbs, or modal verbs. This indicates that the questioner wants to verify the truth of an idea when he or she already has a specific idea.

• *Does the oxygen bonded with another oxygen don't count as another oxidation state?*

• *Are the cranial nerves part of the CNS and the spinal nerves part of the PNS?*

CONSEQUENCE questions In 11.58% of the questions we inspected, the questioners asked for the consequences or results of an event. Most of the questions include *What happen*, *How affect* and so on. This indicates that the questioner is trying to draw a conclusion about the effects or consequences of an action.

• *What happens to the water at the bottom of the container?*

• *How will the viscosity of liquid be affected by increase in temperature?*

According to (Craig et al., 2000), six question categories involve deep reasoning: causal antecedent, causal consequence, goal-orientation, enablement, instrumental/procedural, and expectational. Connecting it to KHANQ, PROCEDURAL questions, CAUSE questions, and CONSEQUENCE questions are three categories that involve deep reasoning, accounting for 51.58%.

4.4 Prompt types

We further analyzed the prompts of these questions and divided the prompts into four major types:

- *Condition* (36.94%): *Prompt* is a conditional clause (trigger word: "if", "when", etc.);
- *Preposition* (8.32%): *Prompt* is a prepositional phrase (trigger word: "in", "for", etc.);

Context:

Passive transport basically does not require any form of energy compared to active transport. In the case of osmosis,the water moves from areas of HIGH WATER concentration to areas of LOW WATER concentration, which makes it a form of passive transport. It uses no energy to move, it just drifts into lower concentrations of WATER. WATER not other materials, only WATER. Osmosis deals with water.

Prompt:		**Question:**
Condition	If water moves from areas where solutes are less concentrated to areas where they are more concentrated	Why would osmosis be a form of Passive Transport?
Preposition	In the case of osmosis	How molecules move from areas of high concentration to low concentration?
Citation	Passive transport is when molecules move from areas of high concentration to low concentration	Shouldn't osmosis technically be classified as a form of active transport?
Question	Are the modes of transport that move molecules from high to low concentrations all passive transport?	Is osmosis also passive transport for water?

Table 3: Different questions raised by learners for the same context with different types of prompts.

- *Citation* (33.95%): *Prompt* is a complete sentence expressing some known information or the learner's own understanding;

- *Question* (20.79%) - *Prompt* is an initial question which leads to the followup question.

Given the same context, different prompts trigger different questions. In Table 3, we show a typical example in KHANQ in which four different types of prompts lead to different questions. We observe a strong correlation between the prompt type and the question type, which reveals how the prompt affects the question: 1) Most of the questions asked under the *condition-type prompt* aim to seek for new information based on the condition set by the questioner. The questions generally ask for causes or consequences. There is a strong logical connection between the question and the prompt, but the question rarely repeats the words in the prompt; 2) Most of the questions asked under the *preposition-type prompt* are about a specific object or phase. They generally ask for procedures or methods. These questions are general and often do not stand alone without the prompt; 3) Most of the questions asked under the *citation-type prompt* are to verify what the questioner already knows or to doubt the newly learned content based on what the learner is already known. They are mainly VERIFICATION questions. The questions tend to repeat the keywords in the prompt; 4) Most of the questions asked under the *question-type prompt* are specifications of the previous question asked by the questioner. The types of the question in *Prompt* and *Question* are usually the same.

5 Experiments

We evaluate the performance of state-of-the-art question generation models on KHANQ, focusing on the following:
• Exploring whether existing models can generate reasonable educational questions by conducting both automatic evaluation (Section 5.3) and human evaluation (Section 5.4).
• Analyzing whether the model needs to actually understand a certain type of prompts when generating questions. (Section 5.5).

5.1 Experimental Settings

We formulate the question generation task as predicting the *Question* given both *Context* and *Prompt* as inputs. Through preliminary experiments, we find both context and prompt are necessary to form a well-defined QG setting because 1) if the prompt is not given, many possible questions can be asked for the context paragraph. The model does not have any guidance on what to ask, leading to simple trivial questions in most cases; 2) if the context is not given, the problem becomes a simple language modeling task which aims to generate the missing part of an incomplete question. In our experiment, 90% of the data in KHANQ are used for training while the remaining are used for testing.

5.2 Models

We choose five generation models based on pretraining language models (PLMs) that perform well on QG tasks for evaluation. We also tried QG-specific models without pretraining such as Info-HCVAE (Lee et al., 2020), but we found that they fail to generate meaningful questions in KHANQ.

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
Human Baseline	31.29	16.74	10.72	7.47	15.60	30.83
BertGeneration (Rothe et al., 2020)	17.83	5.75	2.52	1.15	8.07	17.80
GPT-2 (Radford et al., 2019)	17.01	5.59	2.29	1.31	7.79	18.78
BART (Lewis et al., 2020)	25.10	11.22	6.20	3.41	12.71	26.14
Google-T5 (Raffel et al., 2020)	25.62	12.93	7.25	4.32	12.66	27.62
UniLM (Dong et al., 2019)	20.15	8.83	4.65	2.76	10.76	22.02

Table 4: Automatic evaluation results for different models in KHANQ with BLEU1-4, METEOR and ROUGE-L

	Appro.	Cov.	Coh.	Comp.
Human	**4.32**	**4.44**	**4.13**	**3.83**
BertGeneration	2.63	2.12	2.35	3.47
GPT-2	2.87	2.69	3.07	2.40
BART	4.28	3.65	3.37	3.41
Google-T5	3.90	3.74	3.31	3.75
UniLM	3.35	3.26	2.96	3.37

Table 5: Human evaluation results for different models in KHANQ on appropriateness (Appro.), coverage (Cov.), coherence (Coh.) and complexity (Comp.)

BertGeneration (Rothe et al., 2020) This is a Transformer-based sequence-to-sequence model in which the parameters of both encoder and decoder are initialized with BERT (Devlin et al., 2019). We use the implementation from Huggingface[2].

GPT-2 (Radford et al., 2019) GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. We fine-tune GPT-2 on our training data, by formatting the input sequence as: `Context [PRT] Prompt [QUE] Question`. During test time, `Context [PRT] Prompt [QUE]` is given to predict the question.

BART (Lewis et al., 2020) BART is consists of a bidirectional encoder and a left-to-right decoder. The pretraining task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token. We fine-tune BART on KHANQ by predicting the question given the context and the prompt.

Google-T5 (Raffel et al., 2020) Google-T5 is another state-of-the-art language generation model which pretrains the Transformer with the fill-in-the-blank-style denoising objectives. The model is trained to recover missing words in the input. We use the `t5-base` model provided by Huggingface and fine-tune it on the training data of KHANQ.

UniLM (Dong et al., 2019) UniLM uses three types of language modeling tasks for pretraining: one-way, two-way, and sequence-to-sequence prediction. We initialize UniLM with the parameters of BERT-base (Turc et al., 2019) and then fine-tune it on KHANQ and predict the question.

5.3 Automatic Evaluation

We evaluate the generated questions using BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and ROUGE-L (Lin, 2004).

To compare with human performance, we recruited two high school graduates who were not involved in the annotation process and asked them to perform the same task as models. We took a sample of 80 data and asked them to generate questions based on the context and the prompt. To reduce subjective differences, they were required to develop the results after discussion.

Table 4 shows that BART and Google-T5 perform the best on KHANQ, with similar performance. BART achieves the best METEOR score, while Google-T5 achieves the best BLEU1-4 and ROUGE-L. Both these two models perform significantly better than other models. This observation is consistent with other language generation tasks in which BART and Google-T5 also show strong performance in generation. However, although BART and Google-T5 have achieved super-human performance in generating shallow questions in SQuAD, in our KHANQ dataset, all models perform worse than the human baseline in all metrics, indicating that KHANQ is a more challenging benchmark for SOTA models. In Appendix B, we show examples of generated questions by different models.

[2] `https://huggingface.co/transformers/model_doc/bertgeneration.html`

Model	prompt type	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
BART	condition	+0.06	-0.26	-0.61	-1.17	-0.55	+0.80
	preposition	-2.33	-1.35	-1.06	-0.39	-1.56	-0.92
	citation	-0.56	-0.16	+0.09	+0.63	-0.10	+1.26
	question	-0.60	+1.43	+1.71	+1.92	-0.34	+2.36
Google-T5	condition	-1.13	-1.67	-1.29	-1.04	-1.53	-2.73
	preposition	-4.33	-4.90	-4.23	-2.96	-2.42	-2.87
	citation	-2.52	-2.45	-1.07	-0.36	-0.18	-1.69
	question	-0.36	-0.66	-0.08	+0.16	-0.96	-1.23

Table 6: Changes in the automatic evaluation scores of BART and Google-T5 when data samples from different type of prompts are used as the test set.

5.4 Human Evaluation

We conduct human evaluation on same sets of test samples used in Section 5.3. Each set consists of the human written question and five questions generated by five different models. We hired three annotators who participated in our data annotation to rate the total 480 questions with the best being a 5 and the worst being a 1 on four criteria: (1) *Appropriateness*: whether the question is semantically correct, regardless of the context and the prompt; (2) *Coverage*: whether the question is derived from the context and covers most of the information in the context; (3) *Coherent*: whether the question is coherent with the prompt; (4) *Complexity*: whether the question involves deep reasoning. We further asked them to give what the best and worst scores should be for each aspect to adjust the absolute difference between the best and worst outputs.

As shown in Table 5, BART and Google-T5 also perform better than other models in human evaluation, which is consistent with the automatic evaluation. Human reference still achieve the highest scores on all four aspects, indicating that QG models still fail to reach the human level. BART generates questions with the best appropriateness, and Google-T5 generates questions with the best complexity. In both aspects, they are very close to the human baseline. This suggests that BART and Google-T5 have the ability to ask fluent and complex questions similar to humans. However, the scores of coverage and coherence have a large gap with the human reference for all models, indicating that although the generated questions look fluent by themselves, they often do not cover the essential knowledge covered in the context or they fail to be coherent with the given prompt. This shows that although pretraining models are powerful in generating fluent-looking questions, generating questions that require a deep understanding of the context and the prompt is still challenging.

5.5 Analysis of the effects of prompt

In this section, we analyze whether the model needs to actually understand a certain type of prompts when generating questions. We take turns using data samples from one type of prompts as the test set and the other three as the training set. We choose to analyze BART and Google-T5, which perform well in the previous evaluations.

As shown in Table 6, when the training and testing sets are divided by prompt type instead of randomly, the generation effect of Google-T5 will be much worse, while the generation effect of BART will not change much, which indicates that BART has a better transfer learning ability under different prompt types than Google-T5. It also shows that the questions under different prompt types have some commonality. It is worth noting that if the training set does not contain data samples with the *preposition-type prompt*, both models perform worse when generating questions under this type of prompt. This can be attributed to the fact that *preposition-type prompt* tends to contain very limited amount of information, and most of the questions under those prompts are in-depth questions asking about processes or methods. Generating such questions requires a very accurate understanding of the prompts, which cannot be achieved by a model that has not been trained with these samples.

6 Conclusions and Future Works

In this paper, we propose KHANQ, a dataset for generating in-depth educational questions. Each sample in KHANQ is carefully annotated as *Context*, *Prompt*, and *Question* to form a clean dataset. We evaluate the performance of state-of-the-art question generation models on KHANQ. We find that although it is feasible for the model to generate fluent and complex questions, the ability to understand and reason over the context and the prompt is still far from reaching the human level.

There are several future directions that are worth investigating: (1) what different results the models will produce in terms of different types of questions; (2) how to enable models to obtain information from the context and the prompt separately and then make the inference, to enhance their ability to seek information under the given conditions.

References

Seohyun Back, Akhil Kedia, Sai Chetan Chinthakindi, Haejun Lee, and Jaegul Choo. 2021. Learning to generate questions by learning to recover answer-containing sentences. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP)*, pages 1516–1529.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *ACL/IJCNLP (1)*, pages 6424–6439.

Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. In *ICWSM*.

S. D. Craig, B. Gholson, M. Ventura, and A. C. Graesser. 2000. Overhearing dialogues and monologues in virtual tutoring sessions: Effects on quesioning and vicarious learning. *International Journal of Artificial Intelligence in Education*, 11:242–253.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *CoRR*, abs/1905.03197.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 866–874.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. *CoRR*, abs/1908.04364.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Kalpesh Krishna and Mohit Iyyer. 2019. Generating question-answer hierarchies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2321–2334, Florence, Italy. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jose Otero and Arthur C Graesser. 2001. Preg: Elements of a model of question asking. *Cognition and instruction*, 19(2):143–175.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised multi-hop question answering by question generation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 5866–5880.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1463–1475.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.

Luu Anh Tuan, Darsh J. Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring question-specific rewards for generating deep questions. In *International Conference on Computational Linguistics (COLING)*, pages 2534–2546.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Jianxing Yu, Xiaojun Quan, Qinliang Su, and Jian Yin. 2020. Generating multi-hop reasoning questions to improve machine reading comprehension. In *International World Wide Web Conference (WWW)*, pages 281–291.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3901–3910.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. 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 6032–6037, Hong Kong, China. Association for Computational Linguistics.

A Data Annotation Procedure

A.1 Filter Questions

Purpose: Questions should: 1. be deep enough; 2. contain no external information that requires a deep query; 3. be able to be rewritten as prompt + question.

Filtering out:

- **Question cannot be rewritten as prompt + question**
  Question: what about names from 16-19?

- **Questions that require mathematical calculations but do not provide formulas** (trigger word: number, units)
  Question: a car travels the first half distance between two places at 40kmph and the other half at 60kmph. what will be the average speed of the car?

- **Questions that cite unavailable timestamp or figure without specifying the corresponding text** (trigger word: "at 0:40", "in figure 1 below")
  Question: When the instructor refers to "Lesser Apes" at 0:40, what characteristics classify these "Lesser Apes"?

- **Simple questions that ask for definitions** (trigger word: "What is")
  Question: what is fasciculus

- **Questions that are not understood by the answerer** (trigger word: "Do you mean" in the answer)

Question: how do i know whether specific molecules will undergo active or passive diffusion by just looking at the molecule?
Answer: There is no 'active' diffusion. Diffusion is passive transport. Do you mean diffusion or facilitated diffusion? It depends what 'liquid' is moving through your semipermeable membrane.

A.2 Rewrite Question into Prompt + Question

Purpose: Generate Prompt and Question
* Highlighted part is Prompt, other part is Question

- **The original question is a conditional clause + question, and the conditional clause can be used as prompt** (trigger word: "if", "when")
  If there is bacteria in our blood and there is only 1% of white blood cells, wouldn't that take a long time to dispose of the bacteria?

- **The original question is prepositional phrase + question, and prepositional phrase can be used as prompt** (trigger word: "in", "on", "for")
  Does binary fission occur in the same way for ALL bacteria?

- **A sentence that quotes a part of the article can be used as a prompt** (trigger word: "in …th paragraph", "in …th section")
  In the first section you mention a graph of cyclin levels over the expression cycle throughout mitosis. Why is G1 Cyclin required throughout the entire cyclin expression cycle of mitosis?

- **Questions that cite their own knowledge, the cited knowledge can be used as prompt** (trigger word: "I understand", "I thought", "I know")
  It took me a while to figure out that DNA isn't just one molecule, but a collection of molecules, one per chromosome in humans. Why do people still call DNA a molecule?

- **The question consists of multiple sentences, using the preceding declarative sentence as the prompt** (trigger word: "." ",")
  There are four phases in one cell cycle (G1, S, G2 and M). Apoptosis occurs in which phase?

Special cases:

- **If there are multiple consecutive questions** (trigger word: ?) **First generate Prompt and Question for each question according to the above standards, then if the later question is asked on the basis of the earlier question, the earlier question is used as the Prompt of the later question**
  Is autism a genetic disorder? If so, which chromosome determines the mutation?
  ⇒ (The first question cannot be written as prompt + question, filter out. There is no question before, no further processing. The second question cannot be written as prompt + question, filter out. Further processing:)
  Prompt: Is autism a genetic disorder? If so.
  Question: which chromosome determines the mutation?

- **For the question that can generate multiple Prompt, combine multiple Prompts into one Prompt**
  how bonds require energy in order to be broken and vice versa, why is it opposite for ATP bonds? Because when ATP bonds are broken, energy is released
  ⇒
  Prompt: Because when ATP bonds are broken, energy is released, how bonds require energy in order to be broken and vice versa,
  Question: why is it opposite for ATP bonds?

- **For the choice question** (trigger word: "or") **if one option does not contain information, delete it. if the two choices contain different information, split it into two questions**
  During radioactive decay, can the neutron be kicked out of the nucleus? Or is it only the proton which gets kicked out?
  ⇒
  1 Prompt: During radioactive decay, Question: can the neutron be kicked out of the nucleus?

  2 Prompt: During radioactive decay, Question: is it only the proton which gets kicked out?

A.3 Filter Answers

Purpose: Answers should: 1. be able to be rewritten into independent contexts; 2. contains enough information to ask the corresponding questions.

Filtering out:

- **Answers contains too little information**
  Answer: The H zone is the space in the middle of the sarcomere where only myosin proteins are found.

- **Answers that cannot stand alone as a paragraph**
  Answer: Experimentation. It is not determined from an equation. In a similar way to how you cannot solve for the specific heat of a substance, you can conduct and experiment to find it, or use an accepted value from a table.

- **Answers that cannot answer the question**
  Question: I thought an energy-releasing reaction was called an exothermic reaction and a reaction that takes in energy is endothermic. In the article, it defines them as exergonic and endergonic. Are they the same?
  Answer: Exothermic and endothermic refer to specifically heat. Exergonic and endergonic refer to energy in general.

- **Answers including background knowledge that are not mentioned**
  Answer: Not all ions are reactive (think of dissolving salt in water to give Na+ and Cl- ions) - it depends on the circumstances. H+ ions are more reactive than H3O+ ions, so when an acid dissociates in the water, the protons immediately latch on to water molecules to give H3O+ ions which are more stable than H+ ions.

Special cases:

- **If a usable question has multiple answers, first filter according to the above standards. For the filtered-out answers, first join them together and keep the joined answers if they meet the requirements after joining. Finally, each kept answer can be used as a context for a data**
  Answer: The clouds keep our temperature regulated. If we didn't have clouds, it would be extremely hot at night and extremely cold at night. Aren't you glad God created clouds? He really thought it out well when he created the earth. Hope this helps. (The answer is not good enough, but a good context can be generated after joined)
  Answer: It would be both hot and cold (bit like the moon). When the planet faces the sun, it would be really hot and when it faces away from the sun, it would be really cold. (The answer is not good enough, but a good context can be generated after joined)
  Answer: If you just mean a place hardly has clouds. I think hotter? For there are no raining. Whatever, the climate will become very hard. But if you mean that there wasn't a state called cloud, then I don't know.... What do you actually mean about no clouds? Because clouds are the basic state for water. If there aren't any clouds, what will water become? (The answer is not good enough, contains too much information that is invalid or even contradictory to other answers, and a good context cannot be generated after splicing, so discard it directly)
  ⇒
  Answer: The clouds keep our temperature regulated. If we didn't have clouds, it would be extremely hot at night and extremely cold at night. ~~Aren't you glad God created clouds? He really thought it out well when he created the earth. Hope this helps. Answer:~~ It would be both hot and cold (bit like the moon). When the planet faces the sun, It would be really hot and when it faces away from the sun, It would be really cold.

A.4 Rewrite Answers into Contexts

Purpose: Generate Contexts
*Before deletion is the answer, after deletion is the context

- **Remove answer-tone phrases for the question** (trigger word: "yes", "no", "short answer:" , "this is because")
  Answer: ~~Short answer:~~ a photon is a particle of light. ~~Longer answer:~~ light is energy. Sometimes we think of light as being a wave in the form of an electro-magnetic wave but other times it can be described as a particle. A photon in this case, is 1 unit of light with a variable amount of energy which depends on its frequency.

- **Remove external information from the subject irrelevant(need for deep query) content** (trigger word: Url, email, phone, etc)
  *Answer: *Oncogene* are mutated genes that switch from G1 to S phase, (even when there is no need for cell division). So they are accečerators of the process. You know*

*that is a checkpoint of a cell. However, there are also *tumor suppressor* genes and they act a brake. So whilst the tumor suppressor is sitting there in the cell, it's stopping the cell from going around the cell cycle. If we, again, trigger the cell cycle or attempt to by instructing the cell with a signal initiating the relay, one of the jobs of that relay is to remove that block, that brake. As long as a tumor suppressor is working, cancer will not arise. If you remove tumor suppressor (mutation) the cell is free to move from G1 to S phase. If they are both in mutant form, cancer arises. ~~https://www.ncbi.nlm.nih.gov/books/NBK21662/ https://www.futurelearn.com/courses/inside-cancer/14/steps/579660a~~*

- **Remove questions that were asked but not answered** (trigger word: "?")
  Answer: DNA is DNA it is universal in all organisms. However, combinations of nucleotides (codons) are different and code for different amino acids. ~~Why do you think it should be tested in other organisms?! Sequencing whole genomes have already been done. If it is not enough, what it is?~~

A.5 Paraphrasing

Purpose: 1. Delete words that do not contain information; 2. Rewrite to formal tone.

- **Fix grammar and spelling problems**
  Mathematically, sphere or ~~an~~a circle has more area compared to other geometric shapes. so, ~~we cant consider~~can we consider neopentane as spherical?*

- **Delete daily terms** (trigger word: "okay")
  ~~Okay,~~where do the single protons, the hydrogens come from? How do we add them to our equation?

- **Delete the conversational language between the questioner and the answerer** (trigger word: "Hello", "good question", "hope it helps", "sincerely", "remember")
  ~~Remember,~~ velocity and acceleration are vector quantities, which have both magnitude and direction (+/-). ~~Hope that helps~~

- **Delete the conjunctions that appear at the beginning of the sentence** (trigger word: "So", "But")
  ~~so~~ we evolve from apes, so if one mated with a human could you see the "half-way" of human evolution?

- **Delete subjective prompt words** (trigger word: "I believe")
  ~~I thought~~ red blood cells are already larger in diameter than the smallest capillaries, so that they have to squeeze through.

- **Delete quote cues and quotation marks** (trigger word: "From the video, it says that")
  ~~In the third paragraph under "Ionic Bonds", it says that~~ there is no such thing as a single NaCl molecule. Why can't you have a single molecule of NaCl?

- **Delete timestamp** (trigger word: "At 3:40")
  ~~At 3:40, the speaker says that~~ enthalpy of a reaction can be calculated using Hess's law or by using the Delta H of formation. I thought Hess's law uses the Delta H of formation, but she seems to be saying those are two different ways. is there a difference between using Hess's law and using the Delta H of formation, and if so, what is it?

- **Revise the personal pronouns referring to the questioner or answerer** (trigger word: "you", "I")
  How ~~can I~~to figure out the extent of branching for a molecule?

- **If the specific noun referred to by the pronoun is in another part, change the pronoun to the specific noun it refers to** (trigger word: "it", "they", "this", "that")
  Condition: isn't there a sulfur cycle as well as all the others and if so
  Question: how does ~~it~~a sulfur cycle as well as all the others work

- **Rewrite the symbols that will be recognized as garbled** (trigger word: "°C")
  Having its boiling point at ~~°C~~ degree Celsius?

A.6 Mark source

Purpose: source should: 1. contains enough information to identify the origin of the data; 2. unaffected by website URL changes.

Use the title of the text as the source

B Case study

Context:
Superfluid means the substance is at the point between liquid and gas (equilibrium)at very high temperature and pressure. helium becomes a superfluid at extremely low temperatures. superfluid is strange in the way that it seems to have a viscosity of zero, it can flow up the sides of a container, and many other creepy things.

Prompt:
helium can become a ’superfluid’ . while helium in its liquid state

Question:
Human ⇒ Under what conditions does helium become superfluid ?

BertGeneration ⇒ why does hydrogen *need to dissolve in water*?
GPT-2 ⇒ helium can never make a superfluid . Does the *helium in its liquid state always have viscosity of zero*?
BART ⇒ Why does helium become a ’superfluid’ ?
Google-T5 ⇒ How can helium become a superfluid ?
UniLM ⇒ how can helium become a superfluid ?

Context:
The efficiency of the respiratory system is very very efficient. Imagine a system of pipes; one pulls the air , another transfers it. Since pipes are rounded and therefore ’closed’ air cannot escape through the piping. Take for example your hand. Curl it, and blow through it. Almost none or if any air escapes, except through the other side. Because of this, the only path for the air to take is to the lungs. In which the lungs work/use the air and the entire cycle happens again. When a biological system is working many other dependent systems are working too.

Prompt:
As the air keeps moving down the throat to the lungs, does some air leak or escape?

Question:
Human ⇒ A. Where is the path of air passage through breathing space and where the working place is?

BertGeneration ⇒ Does something like *a hydroxylation* change the oxygen species?
GPT-2 ⇒ how does the oxygen diffuse in or would it go in or can the same pressure?
BART ⇒ Why does the respiratory system work?
Google-T5 ⇒ Why is the respiratory system so efficient?
UniLM ⇒ how does the air move in a system ?

Table 7: Examples of questions generated on KHANQ: 1) States of matter 2) The lungs and pulmonary system. Colored Text indicates key words and *Italics* indicate inappropriate words