RESEARCH-ARTICLE

Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models

SUBHANKAR MAITY, Indian Institute of Technology Kharagpur, Kharagpur, WB, India

ANIKET DEROY, Indian Institute of Technology Kharagpur, Kharagpur, WB, India

SUDESHNA SARKAR, Indian Institute of Technology Kharagpur, Kharagpur, WB, India

Published: 15 December 2023

Citation in BibTeX format

FIRE 2023: Forum for Information Retrieval Evaluation
December 15 - 18, 2023
Panjim, India

FIRE '23: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (December 2023)
https://doi.org/10.1145/3632754.3632755
ISBN: 9798400716324

Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models

Subhankar Maity*
IIT Kharagpur
West Bengal, India
subhankar.ai@kgpian.iitkgp.ac.in

Aniket Deroy
IIT Kharagpur
West Bengal, India
roydanik18@kgpian.iitkgp.ac.in

Sudeshna Sarkar
IIT Kharagpur
West Bengal, India
sudeshna@cse.iitkgp.ac.in

ABSTRACT

Designing high-quality educational questions is a challenging and time-consuming task. In this work, we propose a novel approach that utilizes prompt-based techniques to generate descriptive and reasoning-based questions. However, current question-answering (QA) datasets are inadequate for conducting our experiments on prompt-based question generation (QG) in an educational setting. Therefore, we curate a new QG dataset called *EduProbe* for school-level subjects, by leveraging the rich content of NCERT textbooks. We carefully annotate this dataset as quadruples of 1) *Context*: a segment upon which the question is formed; 2) *Long Prompt*: a long textual cue for the question (*i.e.*, a longer sequence of words or phrases, covering the main theme of the *context*); 3) *Short Prompt*: a short textual cue for the question (*i.e.*, a condensed representation of the key information or focus of the *context*); 4) *Question*: a deep question that aligns with the *context* and is coherent with the *prompts*. We investigate several prompt-based QG methods by fine-tuning pre-trained transformer-based large language models (LLMs), namely PEGASUS, T5, MBART, and BART. Moreover, we explore the performance of two general-purpose pre-trained LLMs such as Text-Davinci-003 and GPT-3.5-Turbo without any further training. By performing automatic evaluation, we show that T5 (*with long prompt*) outperforms all other models, but still falls short of the human baseline. Under human evaluation criteria, Text-Davinci-003 usually shows better results than other models under various prompt settings. Even in the case of human evaluation criteria, QG models mostly fall short of the human baseline. Our code and dataset are available at: https://github.com/my625/PromptQG

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; *Language resources*; • **Applied computing → Education**.

KEYWORDS

Education, Question Generation, Prompt, Large Language Models (LLMs)

*Corresponding Author

FIRE 2023, December 15–18, 2023, Panjim, India

ACM ISBN 979-8-4007-1632-4/23/12...$15.00
https://doi.org/10.1145/3632754.3632755

ACM Reference Format:
Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2023. Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models. In *Forum for Information Retrieval Evaluation (FIRE 2023), December 15–18, 2023, Panjim, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3632754.3632755

1 INTRODUCTION

The primary objective of the automated question generation task (AQG) is to automatically produce questions based on textual or knowledge data. Prompt-based QG refers to the approach of generating questions using a prompt or stimulus text, which enables providing more information while producing questions [8, 15].

Previous studies in AQG [11, 30, 35] employ single-hop question-answering (QA) datasets such as SQuAD [27], which are representative of QA research, as well as multi-hop QA datasets such as HotpotQA [32]. But they were considered unsuitable for this particular study, which specifically focuses on educational materials like textbooks for real-life classroom scenarios.

Our proposed dataset *EduProbe* is sourced from NCERT [1] textbooks and covers a wide range of subjects and grade levels from 6^{th} to 12^{th} standard. We carefully annotate this dataset as quadruples of 1) *Context*: a segment upon which the question is formed; 2) *Long Prompt*: a long textual cue for the question (*i.e.*, a longer sequence of words or phrases, covering the main theme of the *context*); 3) *Short Prompt*: a short textual cue for the question (*i.e.*, a condensed representation of the key information or focus of the *context*). Primarily, we select a noun phrase from the beginning portion of the *context* to serve as a *short prompt*; 4) *Question*: a deep question that aligns with the *context* and is coherent with the *prompts*. After annotation, we gather a total of 3,502 <*Context*, *Long Prompt*, *Short Prompt*, *Question*> quadruples to form the *EduProbe* dataset. An example is given in Table 1.

Table 1: One instance within our *EduProbe* dataset.

Context: Purchasing power parity (PPP) is an economic indicator that signifies the purchasing power of the currencies of various nations of the world against each other. It helps in comparing living standards between different countries and estimating economic productivity.
Long Prompt: purchasing power parity helps
Short Prompt: purchasing power
Gold Standard Question : What does purchasing power parity do?

To go into the diversity and depth of questions in our created dataset, we classify questions based on the first two words in the question and make comparisons to other frequently used QA datasets, such as SQuAD, and HotpotQA. The corresponding table

[1] https://en.wikipedia.org/wiki/National_Council_of_Educational_Research_and_Training

is provided as Table 2. As stated in [6], questions phrased with "*Why is*" and "*How do*" are often considered to be deep questions. Table 2 shows that our proposed dataset, *EduProbe* consists of more deep questions (starting with "*Why is*", and "*How do*") compared to SQuAD, and HotpotQA. Following the criteria in [3], we further categorize these questions based on their reasoning type. It turns out that 46.14% of questions in *EduProbe* include deep reasoning.

Table 2: Leading bigrams that occur most frequently in *EduProbe*, SQuAD, and HotpotQA.

EduProbe	**%**	**SQuAD**	**%**	**HotpotQA**	**%**
What is	17.70	What is	8.5	What is	5.0
What are	15.10	What was	5.3	Who was	2.1
Why is	7.31	How many	4.9	What was	2.0
What was	2.94	When did	3.1	In what	1.8
Which is	2.79	In what	2.9	When was	1.7
How many	2.34	What did	2.8	Who is	1.6
How do	2.08	When was	2.1	How many	1.0
Who was	1.91	Who was	2.1	In which	0.9
Who is	1.91	What does	1.7	What year	0.9
What do	1.88	What are	1.7	Are both	0.9

In this work, we fine-tune four transformer-based large language models (LLMs) such as Pegasus [33], T5 [26], MBART [18], and BART [16] on our proposed dataset *EduProbe*. With the recent advancement of general-purpose LLMs[2] such as Text-Davinci-003 and GPT-3.5-Turbo, the question arises of whether these general-purpose LLMs can be used for QG without any further training. So we also explore pre-trained LLMs such as Text-Davinci-003 and GPT-3.5-Turbo for prompt-based QG. Through a comprehensive evaluation, we observe that T5 (*with long prompt*) outperforms other models in all automated evaluation metrics (see Section 8). In the case of automated metrics, the QG models still fall short of the human baseline. Regarding human evaluation (see Section 9), Text-Davinci-003 (*with long prompt*) shows the best proficiency in generating questions in terms of *grammaticality*. Furthermore, Text-Davinci-003 (*with short prompt*) demonstrates commendable proficiency in generating questions in terms of *appropriateness*, and *relevance*. On the other hand, Text-Davinci-003 (*without prompt*) distinguishes itself by generating questions that are *novel* and *complex*. Although in the case of manual evaluation criteria, the QG models still fall short of the human baseline except for Text-Davinci-003 and GPT-3.5-Turbo under *novelty* and *complexity* criteria.

In summary, the contributions of this paper are as follows: (i) Developing a dataset called *EduProbe*, for school-level subjects, namely History, Geography, Economics, Environmental Studies, and Science, which to the best of our knowledge is not present; (ii) We perform an in-depth comparative study of prompt-based (*e.g.*, *long* and *short prompt*) techniques, and *without prompt*-based technique utilizing state-of-the-art (SOTA) LLMs on our proposed dataset, *EduProbe*. We evaluate them using automated metrics; (iii) Since automated metrics have their own limitations, in terms of evaluating deep questions, we also perform the human evaluation of the generated questions with the help of school-level students and teachers, thereby drawing various meaningful insights.

The remainder of the paper is organized as follows. We discuss the relevant literature in Section 2. We present the motivation of our work in Section 3. We then define the task in Section 4, and discuss the dataset in Section 5. We describe the methodology in Section 6, and the experimental setting in Section 7. We present the automated and human evaluation metrics, results in Section 8, and Section 9 respectively. We have a general discussion on the analysis of the results in Section 10. Finally, we conclude our work in Section 11.

[2]Details of the two LLMs are available at: https://platform.openai.com/docs/models/

2 RELATED WORK

Previous works of QG utilize sequence-to-sequence (Seq2Seq) models [11, 35], to produce questions based on various aspects of the sentence, including its focus, type, and specific relationships. A model is proposed by Pan et al. [22] which is made up of four components: a document encoder that encodes the input document, a semantic graph encoder that embeds the document-level semantic graph using a gated graph neural network based on attention, a content selector that identifies important information from the semantic graph suitable for generating questions, and a question decoder that generates questions based on the enhanced document representation. Xie et al. [31] introduce a framework consisting of two main parts: the question generator and QG-specific rewards. The question generator utilizes a Seq2Seq framework with attention, copying, and coverage mechanisms, similar to existing neural QG works. During training, the model learns by maximizing the likelihood of correct questions. However, this basic question generator faces a problem called exposure bias. To address this issue, they introduce three QG-specific rewards to assess the quality of the questions generated by the basic model. These rewards focus on assessing the *fluency*, *relevance*, and *answerability* of the questions.

2.1 Prompt-based QG

Current works [8, 15] explore a few prompt-based techniques for QG. Gong et al. [8] build a large dataset called KHANQ by annotating each data sample as a triple of <*Context*, *Prompt*, *Question*> and explore prompt-based QG with LLMs such as BERT Generation [29], BART [16], GPT2 [25], T5 [26], and UniLM [7]. The prompts used in KHANQ have been designed on the basis of the learner's background knowledge and understanding of the subject. Prompt-based fine-tuning is employed by Lee and Lee [15] to create multi-hop questions. The methodology for this task involves a series of tasks starting with QG, followed by QA, which is performed repeatedly in cycles to develop a robust methodology for the QG task. They use T5 to train both the QG and the QA models. Also, question paraphrasing is being performed, which adds to the robustness of the method. Finally, prompt-based fine-tuning is performed to generate quality questions. They generate a prompt by selecting relevant words associated with the correct answer.

2.2 Use of LLMs for QG

Here, we discuss the following LLMs, which we explore for the task of QG in this study:

Text-Davinci-003 (abbreviated as Davinci) is a model by OpenAI which has 175 billion parameters. During the training process, a combination of supervised and unsupervised learning methods is used. The specific details of the training data sources used for Davinci have not been publicly disclosed by OpenAI. However, training data are known to consist of a diverse range of sources, including web pages, books, scientific articles, and various other forms of human-written text. The maximum input length supported by Davinci is 4,097 tokens. Davinci is available at: https://platform.openai.com/docs/models/gpt-3-5.

GPT-3.5-Turbo (abbreviated as ChatGPT), an extension of the GPT-3 architecture, which has around 154 billion parameters and underwent training using various text sources such as web pages, books, scientific articles, and more. Its training methods included supervised and reinforcement learning techniques. The primary focus of its optimization was to improve speed, performance, and resource efficiency. It has a maximum input token limit of 4,096. ChatGPT is available at: https://platform.openai.com/docs/models/gpt-3-5.

T5 [26] is based on the transformer architecture and trained using a large-scale dataset consisting of diverse text sources. T5-large has a huge number of parameters, specifically around 770 million, enabling it to capture intricate patterns and relationships in text data. The maximum input length supported by T5-large is 512 tokens. The extensive pre-training and fine-tuning process of T5 makes it a powerful tool for generating high-quality questions and producing accurate natural language output. Here, we use `T5-large` (https://huggingface.co/t5-large) for the QG task.

Pegasus [33] is a transformer-based model designed for text summarization, using an encoder-decoder architecture with self-attention mechanisms to capture long-range dependencies. With approximately 568 million parameters, Pegasus-large can handle complex summarization tasks and generate high-quality summaries. Its input limit is 1,024 tokens. It has the potential to be utilized for the QG task. Here, we utilize `Pegasus-large` (https://huggingface.co/google/pegasus-large) for the purpose of QG.

MBART [18] (Multilingual Bidirectional Auto-Regressive Transformers) incorporates a transformer-based architecture, which allows it to effectively capture contextual information and generate high-quality translations. It leverages pre-training on a large multilingual corpus to learn cross-lingual representations. In terms of the number of parameters, MBART-large-50 has around 610 million parameters. It has a maximum input token limit of 1,024. It can be utilized for the QG task. Here, we utilize `MBART-large` (https://huggingface.co/facebook/MBART-large-50) for QG task.

BART [16](Bidirectional Auto-Regressive Transformers) comprises a bidirectional encoder and a left-to-right decoder. During pre-training, it shuffles the order of sentences and uses a unique method of infilling where sections of text are replaced with a mask token. BART-large has around 406 million parameters. The maximum input length supported by BART is 1,024 tokens. It can be used for QG tasks. Here, we use `BART-large` (https://huggingface.co/facebook/BART-large) for the QG task.

2.3 Datasets used for QG

Due to the data-centric nature of QG, the QG methods mentioned above leverage the availability of large-scale QA datasets, such as SQuAD, HotpotQA, TriviaQA [9], Natural Questions corpus [12], QuAC [5], OpenBookQA [19], etc. According to Cao and Wang [3], these corpora are limited to the generation of simple fact-based questions. Furthermore, as stated in [20, 21], the majority of these QA datasets are borrowed or crowd-sourced from open-source platforms such as Wikipedia articles, and the questions generally do not incorporate multiple sentences as their basis. There is a notable QG dataset for educational purposes called LearningQ [4], which utilizes complete articles or videos as contexts, resulting in a substantial portion of sentences within the contexts being irrelevant to the specific target question. In contrast, we utilize explanatory answers that contain comprehensive knowledge points relevant to the question.

3 MOTIVATION

The creation of high-quality questions is a fundamental task for educators seeking to foster deep understanding and critical thinking in students. However, the process of designing educational questions manually is often burdensome and time-consuming [8]. Through this research effort, our aim is to provide a valuable tool that empowers educators to create descriptive and reasoning-based questions more efficiently. By allowing teachers to allocate more time to classroom interactions and student participation, our proposed QG method has the potential to positively impact teaching practices and enhance learning outcomes.

Previous research in QG [11, 30, 35] predominantly emphasizes the generation of fact-based questions that relate to a single piece of information derived from a single sentence. Moreover, current QA datasets such as SQuAD, HotpotQA, TriviaQA, QuAC, OpenBookQA, etc. do not align with the requirements of generating school-level educational questions since they do not have deep questions that are reasoning-based and descriptive in nature.

So, we curate a new dataset called *EduProbe* to tackle the complexities of school-level educational QG. The proposed dataset serves as the foundation for our investigation into the efficacy of prompt-based methods for QG, which involves providing the system with explicit hints or triggers to generate deep questions. The prompt-based AQG process offered by our approach not only reduces the time required for question development, but also improves the overall quality and variety of questions generated.

Our work has the following three key differences from previous works on QG: (i) Our created dataset *EduProbe* is geared towards creating questions that are more educationally oriented in the context of school-level subjects; (ii) We explore different types of prompt-based techniques (*e.g.*, *long prompt*, *short prompt*, and *without prompt*) with SOTA LLMs (*e.g.*, Text-Davinci-003, GPT-3.5-Turbo, etc.) to provide the QG models additional guidance on what information to emphasize more when generating questions that still have not been explored earlier in a detailed manner; (iii) Our proposed prompt-based approaches are capable of generating a wide variety of questions from a single context which has not been explored in previous works.

4 TASK DEFINITION

In this section, we define three different prompt settings explored in our study:

- **With Long Prompt**: Given a dataset D, where each data point $d \in D$ is represented as a 4-tuple <*Context*, *Long Prompt*, *Short Prompt*, *Question*>, our task is to learn a probabilistic model *P(Question | Context, Long Prompt)* that can generate a relevant question in the context of the given information.

- **With Short Prompt**: Given a dataset D, where each data point $d \in D$ is represented as a 4-tuple <*Context*, *Long Prompt*, *Short Prompt*, *Question*>, our task is to learn a probabilistic model *P(Question | Context, Short Prompt)* that can generate a relevant question in the context of the given information.
- **Without Prompt**: Given a dataset D, where each data point $d \in D$ is represented as a 4-tuple <*Context*, *Long Prompt*, *Short Prompt*, *Question*>, our task is to learn a probabilistic model *P(Question | Context)* that can generate a relevant question in the context of the given information.

Figure 1 illustrates the schematic of the pipeline process used for QG.

5 DATASET

Data Collection and Annotation: There is no public dataset available to conduct our experiments on prompt-based QG in an educational setting. Therefore, we produce a dataset called *EduProbe* by manually creating *question-answer* pairs from segments of varying lengths taken from a diverse set of chapters present in the National Council of Educational Research and Training (NCERT) textbooks on History, Geography, Economics, Environmental Studies, and Science from the 6^{th} standard to the 12^{th} standard.

Firstly, the annotators are instructed to go through the selected chapters line by line and generate *question-answer* pairs by considering only the relevant portions of the information from the NCERT textbooks. To establish the *context*, we instruct the annotators to review the generated answers. Upon analysis, we observe that the majority of the answers consist of comprehensive explanations related to the questions' pertinent knowledge points. Hence, these answers prove suitable for serving as the *context* for the question. Secondly, we instruct the annotators to select a sequence of words or phrases which cover the main theme of the *context* to serve as a *long prompt* for the question. Lastly, annotators are asked to pick up a noun phrase from the beginning portion of the *context* to act as a *short prompt*. In this manner, each *question-answer* pair in *EduProbe* is carefully annotated as a quadruple of <*Context*, *Long Prompt*, *Short Prompt*, *Question*>.

Two annotators (two graduate students) with adequate subject knowledge and experience were involved in manually annotating the data samples.

Data Statistics: We carefully curate 3,502 question-answer pairs, of which 858 pairs are related to History, 861 pairs are related to Geography, 802 pairs are related to Economics, 606 pairs are related to Environmental Studies, and 375 pairs are related to Science. On average, the length of the *context*, *long prompt*, *short prompt*, and question are 55.27 words, 6.80 words, 2.15 words, and 7.16 words, respectively.

Comparing the KHANQ dataset and our proposed dataset *EduProbe* we observe that the average length of the *long prompt* and *short prompt* in our dataset is 6.80 and 2.15 words, respectively. But in the KHANQ dataset, the average *prompt* length is 14.12 words. However, the KHANQ dataset is not publicly available for research purposes.

Question Types: In order to gain a deeper understanding of the question attributes, we conduct a detailed manual analysis of a subset of 65 distinct questions randomly sampled from the *EduProbe* dataset. These questions were categorized according to the criteria specified in [3]. Here, we present a summary of the most commonly found question types in *EduProbe* and their corresponding examples.

- **Procedural Questions**: Out of the questions we examine, 18.46% of them are about the procedures or methods used to achieve a specific outcome. Most of these questions begin with "*How*" and are followed by a modal verb, an auxiliary verb, or "*to*".
  - *How did Pandita Ramabai break stereotypes?*
  - *How did Brahmo Samaj reform Indian society?*
- **Cause Questions**: We observe that 15.38% of the questions examined are focused on the reason or cause behind a concept or event. Most of these questions began with the word "*Why*" and were subsequently followed by a modal verb, an auxiliary verb, or their negative counterparts.
  - *Why is the Ganges river dolphin blind?*
  - *Why is urban waste disposal a serious problem in India?*
- **Verification Questions**: We discover that 9.23% of the examined questions are concerned with verifying the trustworthiness of a concept or event. Most of these questions are formulated as general questions that originate with verbs, modal verbs, or auxiliary verbs.
  - *Does universal basic income (UBI) reduce poverty?*
  - *Are Vedas older than Puranas?*
- **Consequence Questions**: We find that 12.30% of the questions analyzed focused on the consequences or outcomes resulting from a particular event. Most of these questions used phrases such as "*What happens*", "*How does it affect*", etc.
  - *What happens if oceans acidify?*
  - *How does the government deficit affect the economy?*

According to the findings of [3], there are four categories of questions that require profound reasoning, namely cause, consequence, judgemental, and procedural. Relating these categories to *EduProbe*, we observe that the procedural questions, the cause questions, and the consequence questions are three specific categories that involve deep reasoning, representing 46.14%.

6 METHODOLOGY

We experiment with various prompt-based settings with LLMs. There are two main categories of LLMs explored in our study: (1) Pre-trained General-purpose LLMs; (2) Fine-tuned Domain-specific LLMs.

6.1 Pre-trained General-purpose LLMs

We try Text-Davinci-003 (abbreviated as Davinci) and GPT-3.5-Turbo (abbreviated as ChatGPT) model using OpenAI API[3]. For the pre-trained general-purpose LLMs, we have to pass a *prompt* as input to the LLM which will consist of the instruction for the task of QG, along with the *context* and *prompt* (if needed). The LLM will generate text based on the given prompt, which will be our desired

[3]https://platform.openai.com/docs/api-reference/completions

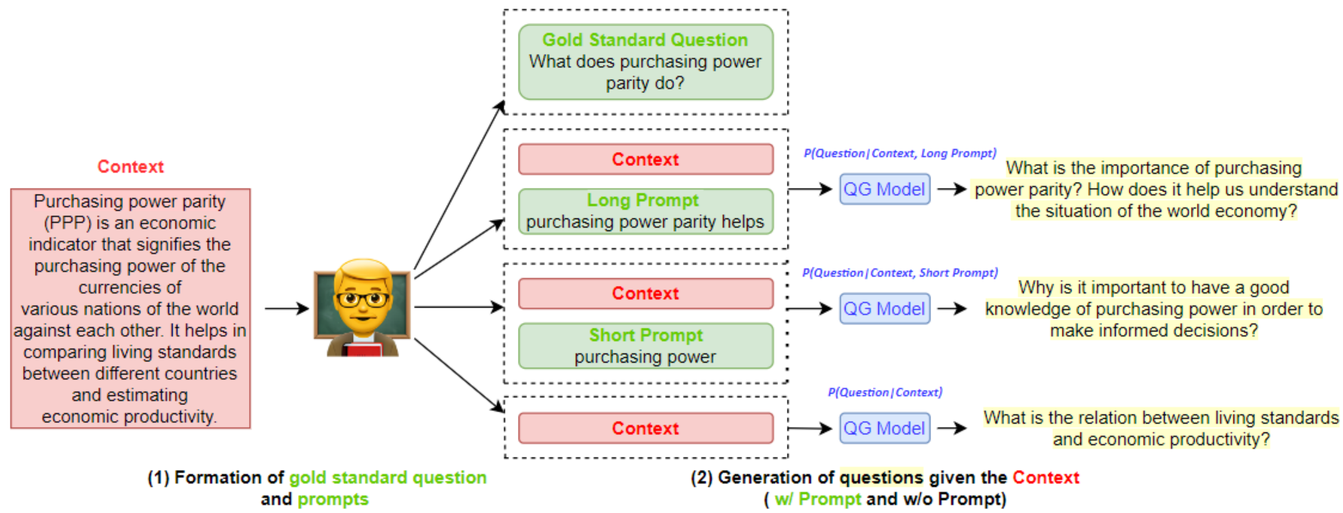

Figure 1: A diagrammatic representation of the pipeline process utilized to generate questions.

output. For each model, we try three different variations based on prompts that are as follows:
a) **With Long Prompt**: The prompt we use is "*Given the context* `<Context>` *and the long prompt* `<Long Prompt>`, *generate a* `Question`".
b) **With Short Prompt**: The prompt we apply is "*Given the context* `<Context>` *and the short prompt* `<Short Prompt>`, *generate a* `Question`".
c) **Without Prompt**: The prompt we utilize is "*Given the context* `<Context>`, *generate a* `Question`".

6.2 Fine-tuned Domain-specific LLMs

We also fine-tune four pre-trained transformer-based QG models (or LLMs), namely Pegasus [33], T5 [26], MBART [18], and BART [16] obtained from the open-source library[4] named Huggingface, in our proposed dataset *EduProbe*. For every model, we try the three different prompt-based techniques which are as follows:
a) **With Long Prompt**: During fine-tuning, we format the input sequence as:(*[CLS]* `Context` *[SEP]* `Long Prompt` *[SEP]*, *[CLS]* `Question` *[SEP]*) pair. We provide *[CLS]* `Context` *[SEP]* `Long Prompt` *[SEP]* as input to predict the `Question` during test time, as shown in Figure 2.

Figure 2: With Long Prompt QG Model.

b) **With Short Prompt**: During fine-tuning, we format the input sequence as:(*[CLS]* `Context` *[SEP]* `Short Prompt` *[SEP]*, *[CLS]* `Question` *[SEP]*) pair. We provide *[CLS]* `Context` *[SEP]* `Short Prompt` *[SEP]* as input to predict the `Question` during test time, as shown in Figure 3.

[4]https://huggingface.co/models

Figure 3: With Short Prompt QG Model.

c) **Without Prompt**: During fine-tuning, we format the input sequence as:(*[CLS]* `Context` *[SEP]*, *[CLS]* `Question` *[SEP]*) pair. We provide *[CLS]* `Context` *[SEP]* as input to predict the `Question` during test time, as shown in Figure 4.

Figure 4: Without Prompt QG Model.

7 EXPERIMENTAL SETTINGS

In our experiment, we randomly sample 80% of the data in *EduProbe* for training and the rest for testing. The experiments are run on an NVIDIA Tesla P100 16GB GPU and the models are optimized using the Adam optimizer [10]. The specific hyperparameter configurations of the LLMs used in our experiments are given in Table 3.

Table 3: Hyperparameters of the LLMs used in our work.

Model	Hyperparameters
Davinci	max tokens: 50, presence penalty: 1.0, frequency penalty: 0.0, temperature: 0.7
ChatGPT	max tokens: 50, temperature: 0.7
Pegasus	learning rate: 2e-3, epochs: 6, batch size: 1, input length: 512
T5	learning rate: 2e-3, epochs: 6, batch size: 2, input length: 512
MBART	learning rate: 1e-3, epochs: 8, batch size: 1, input length: 512
BART	learning rate: 1e-3, epochs: 8, batch size: 1, input length: 512

8 AUTOMATIC EVALUATION

In this section, we present the main results in *EduProbe*, using the methodology explained in Section 6.

8.1 Evaluation Metrics

We use the following popular metrics that compare a QG model-generated question with the gold standard question:
Rouge [17] (Recall-Oriented Understudy for Gisting Evaluation) is widely utilized as a metric to assess the quality of generated summaries by summarization models. Here, we compute Rouge-2 precision, Rouge-2 recall, and Rouge-2 F1 score to evaluate the bi-gram overlap between the QG model-generated questions and the reference gold standard questions. Furthermore, Rouge-L precision, Rouge-L recall, and Rouge-L F1 scores are calculated to measure the longest common subsequence-based match between the generated questions and the gold standard questions.
Meteor[14] (Metric for Evaluation of Translation with Explicit ORdering) calculates the harmonic mean of unigram precision and recall, which is commonly used to evaluate machine translation results. In our case, we apply this metric to measure the unigram overlap between a QG model-generated question and the reference gold standard question.
CHrF[24] (Character n-gram F-score) is a metric that evaluates the similarity between the generated output and the reference summaries at the character level. Here, CHrF calculates the F-score based on the precision and recall of the matching character n-grams between the QG model-generated questions and the reference gold-standard questions.
BLEU[23](Bilingual Evaluation Understudy) is a widely used metric to evaluate the quality of a machine-generated text. Here, we calculate the overlap between the QG model-generated question and the reference gold question based on n-gram matches. BLEU calculates a score ranging from 0 to 1, with higher scores indicating better quality.
BERTScore[34] is a metric that we utilize to measure the similarity between the generated question and the reference gold question using contextual embeddings from a pre-trained BERT model. It computes a score based on the cosine similarity of the embeddings, capturing both lexical and contextual similarities.

We utilize the implementations of the aforementioned metrics from the SummEval package [5].

In order to compare with human performance, we appoint two high school teachers who were not involved in the annotation process and request them to undertake the same task as the models, generating questions based on the given *context* and *prompt* settings for all data samples from the test set. To minimize subjective variations, they were instructed to collaborate and reach a consensus while formulating their responses.

8.2 Results

We present the results of automated evaluation metrics for different models, in order to investigate the influence of prompts. Table 4 represents the results of automatic evaluation of LLMs under different prompt settings for QG. T5 performs the best across all metrics in the *long prompt* setting. When the *prompt is short*, T5 achieves the highest scores in ROUGE-2 precision and ROUGE-L recall. BART obtains the best results in ROUGE-2 Recall, ROUGE-2 F1, ROUGE-L Precision, ROUGE-L F1, METEOR, CHrF, BLEU, and BERTScore under the *short prompt* setting. Furthermore, T5 outperforms the other models in terms of all metrics, except for BERTScore, in the *without prompt* setting and BART achieves the highest BERTScore.

Compared to other models, T5 and BART exhibit superior performance in automatic evaluation across various prompt settings (*e.g.*, *long prompt*, *short prompt*, and *without prompt*). However, human references consistently achieve the highest scores across all automated metrics and prompt settings. This observation emphasizes the fact that SOTA LLMs still have not reached the level of human performance on our *EduProbe* dataset.

[5]https://github.com/Yale-LILY/SummEval

9 HUMAN EVALUATION

Considering the limitations associated with automated metrics in the field of text generation research [1, 2, 28], we also conduct a human evaluation by appointing two high school teachers and three high school students who were not engaged in the annotation process and the generation of human baseline questions. Every human evaluator was asked to rate a total of 1,800 questions, taking into account six models and three prompt settings. The rating scale used ranged from 1 (worst) to 5 (best) based on five criteria: **Grammaticality**, which measures the grammatical correctness of the generated question, regardless of the context or prompt; **Appropriateness**, which examines the semantic correctness of the question irrespective of the context or prompt; **Relevance**, which measures the degree to which the generated question is pertinent and aligned with the given context or prompt; **Complexity**, which estimates the level of reasoning or cognitive effort required to answer the generated question; **Novelty**, which measures the originality and distinctiveness of the generated question in comparison to the gold standard question for the given context.
Model-wise Evaluation: We report the human evaluation results for different models under different prompt settings on the *EduProbe* dataset in Table 5. Davinci demonstrates superior performance compared to other models in human evaluation in different prompt settings (*e.g.*, *long prompt*, *short prompt*, and *without prompt*) in generating questions that exhibit impressive *grammaticality*, *appropriateness*, *relevance*, *complexity*, and *novelty*. However, human references achieve the highest scores on most human criteria, except for the *novelty* under the *long prompt* setting and *complexity* in the *without prompt* setting. This also suggests that SOTA LLMs still fall short of reaching the level of human performance in most cases. Although Davinci and ChatGPT overtake human-level performance in terms of producing *complex* questions.
Inter-annotator Agreement: In order to assess the level of agreement among the five annotators assigned to each generated question, we use Fleiss's kappa as a metric for inter-annotator agreement. Our calculations yield agreement scores of 0.49, 0.43, 0.44, 0.39, and 0.32 for *grammaticality*, *appropriateness*, *relevance*, *complexity*, and *novelty*, respectively. The kappa values in *grammaticality*, *appropriateness*, *relevance* indicate a moderate agreement [13], while the kappa results for *complexity* and *novelty* indicate a fair level of agreement.

Table 4: Automatic evaluation results for different LLMs in *EduProbe* with ROUGE2-Precision, ROUGE2-Recall, ROUGE2-F1, ROUGEL-Precision, ROUGEL-Recall, ROUGEL-F1, METEOR, CHrF, BLEU, and BERTScore. The highest value for any metric in *long prompt*, *short prompt*, and *without prompt* setting achieved by any model is shown in blue. The highest value for any metric achieved by any model is underlined.

Model	ROUGE-2 Precision	ROUGE-2 Recall	ROUGE-2 F1	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F1	METEOR	CHrF (%)	BLEU (%)	BERTScore
With Long Prompt										
Human Baseline	**0.517**	**0.843**	**0.626**	**0.588**	**0.891**	**0.695**	**0.531**	**76.30**	**46.57**	**0.860**
Pre-trained General-purpose LLMs										
Davinci	0.409	0.726	0.491	0.499	0.812	0.603	0.443	68.23	23.24	0.803
ChatGPT	0.391	0.706	0.476	0.484	0.793	0.592	0.423	66.36	22.44	0.790
Fine-tuned Domain-specific LLMs										
Pegasus	0.329	0.798	0.453	0.413	0.885	0.552	0.411	67.95	27.78	0.770
T5	**0.483**	**0.800**	**0.575**	**0.566**	**0.888**	**0.668**	**0.503**	**74.40**	**42.97**	**0.818**
MBART	0.424	0.750	0.526	0.527	0.860	0.640	0.417	71.05	33.55	0.786
BART	0.460	0.794	0.573	0.548	0.887	0.666	0.443	73.72	36.47	0.809
With Short Prompt										
Human Baseline	**0.324**	**0.579**	**0.418**	**0.468**	**0.758**	**0.563**	**0.377**	**61.52**	**24.92**	**0.778**
Pre-trained General-purpose LLMs										
Davinci	0.263	0.492	0.319	0.429	0.723	0.529	0.313	55.62	20.86	0.739
ChatGPT	0.260	0.486	0.304	0.418	0.720	0.522	0.309	54.83	19.54	0.720
Fine-tuned Domain-specific LLMs										
Pegasus	0.240	0.496	0.312	0.374	0.708	0.477	0.331	54.41	18.54	0.723
T5	**0.301**	0.530	0.368	0.448	**0.742**	0.542	0.341	58.76	21.15	0.742
MBART	0.237	0.464	0.308	0.385	0.680	0.483	0.307	53.56	17.64	0.718
BART	0.300	**0.541**	**0.377**	**0.449**	0.740	**0.549**	**0.346**	**59.52**	**21.82**	**0.756**
Without Prompt										
Human Baseline	**0.323**	**0.532**	**0.390**	**0.466**	**0.723**	**0.553**	**0.355**	**58.23**	**23.49**	**0.758**
Pre-trained General-purpose LLMs										
Davinci	0.273	0.462	0.327	0.413	0.666	0.506	0.283	54.86	20.44	0.729
ChatGPT	0.266	0.442	0.319	0.401	0.645	0.492	0.266	52.83	19.44	0.710
Fine-tuned Domain-specific LLMs										
Pegasus	0.214	0.465	0.280	0.346	0.693	0.449	0.307	51.48	16.24	0.702
T5	**0.306**	**0.501**	**0.368**	**0.455**	**0.706**	**0.539**	**0.322**	**57.03**	**21.59**	0.718
MBART	0.219	0.414	0.281	0.373	0.642	0.464	0.293	50.46	17.34	0.706
BART	0.275	0.477	0.341	0.425	0.688	0.516	0.319	54.96	20.05	**0.742**

Table 5: Human evaluation results for different LLMs in *EduProbe* on grammaticality, appropriateness, relevance, complexity, and novelty. The highest value for any metric in the *long prompt*, *short prompt*, and *without prompt* setting achieved by any model is shown in blue. The highest value for any metric achieved by any model is underlined.

Model	Grammaticality	Appropriateness	Relevance	Complexity	Novelty
With Long Prompt					
Human Baseline	**4.95**	**4.97**	**4.48**	**3.98**	**3.10**
Pre-trained General-purpose LLMs					
Davinci	**4.91**	**4.70**	**4.26**	**3.97**	**3.73**
ChatGPT	4.83	4.51	4.20	3.94	3.56
Fine-tuned Domain-specific LLMs					
Pegasus	4.48	4.37	3.74	3.84	3.14
T5	4.14	4.02	3.93	3.53	2.93
MBART	3.94	3.84	3.54	3.42	3.28
BART	3.97	4.00	3.90	3.76	3.20
With Short Prompt					
Human Baseline	**4.94**	**4.95**	**4.43**	**4.16**	**4.08**
Pre-trained General-purpose LLMs					
Davinci	**4.89**	**4.80**	**4.31**	**4.12**	**3.93**
ChatGPT	4.85	4.76	4.28	4.03	3.81
Fine-tuned Domain-specific LLMs					
Pegasus	4.50	4.48	4.06	3.81	3.44
T5	4.80	4.62	4.20	3.96	3.67
MBART	3.38	3.28	3.08	3.14	3.17
BART	3.78	3.85	3.58	3.69	3.38
Without Prompt					
Human Baseline	**4.93**	**4.96**	**4.29**	**4.13**	**4.25**
Pre-trained General-purpose LLMs					
Davinci	**4.90**	**4.61**	**4.18**	**4.25**	**4.05**
ChatGPT	4.82	4.46	4.15	4.18	3.96
Fine-tuned Domain-specific LLMs					
Pegasus	4.42	4.33	3.73	3.62	3.77
T5	4.65	4.45	3.92	3.57	3.84
MBART	3.22	3.00	2.77	3.01	3.40
BART	3.81	3.98	3.60	3.60	3.72

Table 6: A data sample related to Economics in our *EduProbe* dataset, along with corresponding questions generated by different LLMs using various prompt variations.

Context: Purchasing power parity (PPP) is an economic indicator that signifies the purchasing power of the currencies of various nations of the world against each other. It helps in comparing living standards between different countries and estimating economic productivity. **Long Prompt**: purchasing power parity helps **Short Prompt**: purchasing power **Gold Standard Question** : What does purchasing power parity do?
Davinci (*with long prompt*): How can the concept of purchasing power parity (PPP) be leveraged to not only compare living standards and estimate economic productivity between different countries but also address global income inequality and promote equitable economic development on a global scale? **Davinci (*with short prompt*)**: How can the concept of purchasing power be harnessed to foster international collaboration and create a more equitable global economic system that promotes shared prosperity and uplifts disadvantaged nations? **Davinci (*without prompt*)**: In what ways can the utilization of purchasing power parity (PPP) as an economic indicator be expanded beyond its traditional applications to promote sustainable development, address socio-economic disparities, and foster international cooperation for the betterment of global economic systems?
ChatGPT (*with long prompt*): What does purchasing power parity (PPP) indicate and how does it assist in comparing living standards and estimating economic productivity among different nations? **ChatGPT (*with short prompt*)**: What does purchasing power parity (PPP) represent and how does it facilitate the comparison of living standards and the estimation of economic productivity across different nations? **ChatGPT (*without prompt*)**: How does purchasing power parity (PPP) serve as an economic indicator to measure the purchasing power of currencies across different nations, and what are its implications for comparing living standards and estimating economic productivity?
Pegasus (*With long prompt*): What is the importance of purchasing power parity? How does it help us understand the situation of the world economy? **Pegasus (*With short prompt*)**: Why is it important to have a good knowledge of purchasing power in order to make informed decisions? **Pegasus (*Without prompt*)**: What is the relation between living standards and economic productivity?
T5 (*With long prompt*): How does purchasing power parity help in economics? Read more about how does purchasing power parity help in economics? **T5 (*With short prompt*)**: What is the purpose of purchasing power? **T5 (*Without prompt*)**: What is the importance of a country's development index? How does it help the poor?
MBART (*With long prompt*): What does rupee purchasing power parity do for people living in economics? **MBART (*With short prompt*)**: What is the importance of purchasing power in daily living standards between two countries? **MBART (*Without prompt*)**: Why is it important to read a country-based economy pattern? Give example of poverty.
BART (*With long prompt*): How does purchasing power parity help improve economy parity? and why is it good? **BART (*With short prompt*)**: What is the importance of purchasing power in daily living? Purchasing power is important? **BART (*Without prompt*)**: What is the importance of ease of living standards in simple words, and why is it important?

Table 7: A data sample related to History in our *EduProbe* dataset, along with corresponding questions generated by different LLMs using various prompt variations.

Context: During the medieval period in India, Islamic rulers held significant power, leading to the blending of Indian and Islamic cultures, which can still be observed in the architecture and artwork created at that time. The country was governed and administered by notable rulers who made exceptional contributions in the fields of art, culture, literature, and architecture during this period. **Long Prompt**: the contribution of the medieval period to Indian history **Short Prompt**: medieval period **Gold Standard Question**: What is the contribution of the medieval period to Indian history?
Davinci (*with long prompt*): How did the convergence of Indian and Islamic cultures during the medieval period in India influence not only the realms of art, culture, literature, and architecture but also the broader socio-political fabric and historical trajectory of the nation? **Davinci (*with short prompt*)**: How did the dynamic interplay between Indian and Islamic influences during the medieval period shape not only the artistic and architectural landscape but also the social, religious, and intellectual fabric of India, leaving a lasting impact on its history and identity? **Davinci (*without prompt*)**: How did the cross-cultural exchange between Indian and Islamic civilizations during the medieval period in India result in a unique synthesis of artistic, architectural, literary, and cultural expressions, and what enduring legacies did these contributions leave on the country's cultural heritage?
ChatGPT (*with long prompt*): What notable contributions were made to Indian history during the medieval period in terms of art, culture, literature, and architecture, reflecting the blending of Indian and Islamic influences? **ChatGPT (*with short prompt*)**: What impact did the significant influence of Islamic rulers during the medieval period have on the architecture, artwork, governance, and cultural development in India? **ChatGPT (*without prompt*)**: What were the key cultural and artistic developments that emerged during the medieval period in India under the influence of Islamic rulers, and how did these contributions shape the architecture, artwork, literature, and overall cultural landscape of the time?
Pegasus (*with long prompt*): What is the contribution of the medieval period to Indian history? **Pegasus (*with short prompt*)**: What was the impact of the medieval period on Indian culture, politics and economy? Give examples. **Pegasus (*without prompt*)**: What were the major events and developments that took place in the medieval period in India? Give examples.
T5 (*with long prompt*): What is the contribution of the medieval period to Indian history? Have you ever wondered what is the contribution of the medieval period to Indian history? **T5 (*with short prompt*)**: What were the main events of the medieval period in India that took place during the Islamic period? **T5 (*without prompt*)**: What happened in the medieval period in India, which saw a strong control of Islamic rulers?
MBART (*with long prompt*): What is the contribution of the medieval period to Indian history? What is the contribution of the medieval period? **MBART (*with short prompt*)**: What were the major events in the medieval period in India that took place during the ancient period? **MBART (*without prompt*)**: What were the major events that took place in India during the Second World War (WW2)?
BART (*with long prompt*): What is the contribution of the medieval period to Indian history? and what is its significance in Indian history? **BART (*with short prompt*)**: What was the medieval period in India? and what was its importance for shaping public opinion? **BART (*without prompt*)**: What was the main difference between the medieval period in India and the Chalcolithic period?

10 ANALYSIS

Both Davinci and ChatGPT generate questions that differ from the gold standard questions in terms of character, unigram, bigram, or longest common subsequence-based overlap [14, 17, 23, 24]. Therefore, in terms of automated metrics (see Table 4), these two LLMs cannot show good performance. However, the questions generated by both these general-purpose LLMs (*e.g.*, Davinci and ChatGPT) are of good quality, as shown by the results of the human evaluation criteria in Table 5. Davinci and ChatGPT show superior performance in terms of human criteria like *grammaticality*, *relevance*, *appropriateness*, *complexity*, and *novelty*. However, Davinci and ChatGPT fall short of the human baseline in terms of *grammaticality*, *appropriateness*, and *relevance* under the different prompt settings. But it has been observed that Davinci and ChatGPT can generate *novel* questions that are not present in the gold standard, thus giving them high scores for the *novelty* metric under the *long prompt* setting. Overall, Davinci emerges as the best performer, followed by ChatGPT, based on most human evaluation criteria. Furthermore, it has also been observed that Davinci and ChatGPT can generate *complex* questions in the *without prompt* setting. The amount of cognitive effort required to answer a question generated by Davinci and ChatGPT in the *without prompt* setting is significantly higher than the human baseline. Fine-tuned domain-specific LLMs like T5 and BART show good performance in terms of automated metrics because these LLMs generate questions that are closer to the gold standard questions in terms of character, unigram, bigram, and longest common subsequence matches. However, these fine-tuned domain-specific LLMs fall short of Davinci and ChatGPT in terms of human evaluation criteria.

We observe different sets of results under different prompt settings and a broad and diverse range of questions generated from the same context, thereby showcasing the utility of prompt-based QG techniques. It suggests that prompts definitely help to vary the quality of the generated questions which is observed both in the case of pre-trained general-purpose LLMs (*e.g.*, Davinci and ChatGPT) and fine-tuned domain-specific LLMs (*e.g.*, Pegasus, T5, MBART, and BART).

Table 6 and Table 7 show two data samples from our *EduProbe* test set and the corresponding questions generated by different LLMs under various prompt settings.

11 CONCLUSIONS AND FUTURE WORKS

We introduced *EduProbe*, a dataset to create deep and diverse questions that are more educationally oriented in the context of school-level subjects. We explored different types of prompt-based techniques (*e.g.*, *long prompt*, *short prompt*, and *without prompt*) to provide QG models additional guidance on what information to emphasize more when generating questions. The experiments demonstrate that T5 surpasses other models in all automated metrics. Pre-trained general-purpose LLMs such as Davinci exhibit superior proficiency in generating questions that excel in terms of *grammaticality*, *appropriateness*, *relevance*, *novelty*, and *complexity*. Furthermore, Davinci and ChatGPT surpass the human baseline in terms of generating *complex* questions, though they fall short of the human baseline in terms of generating *grammatical*, *appropriate*, *relevant*, and *novel* questions.

We aim to explore even larger language models (*e.g.*, GPT-4) for QG in the future on our proposed *EduProbe* dataset. Currently, we are creating manual prompts through our annotators that can be replaced by automatic keyphrase or span detection models. Additionally, there is a need to develop better automated metrics for measuring the quality of generated questions, as current metrics cannot fully capture the quality of generated questions. We also plan to fine-tune general-purpose LLMs like Davinci in the future. Although LLMs have shown good performance in generating questions, they are still not able to reach human-level performance in most cases. Therefore, further research is required in this direction.

REFERENCES

[1] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics* 47, 4 (12 2021), 861–889. https://doi.org/10.1162/coli_a_00418 arXiv:https://direct.mit.edu/coli/article-pdf/47/4/861/1979827/coli_a_00418.pdf

[2] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9347–9359. https://doi.org/10.18653/v1/2020.emnlp-main.751

[3] Shuyang Cao and Lu Wang. 2021. Controllable Open-ended Question Generation with A New Question Type Ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6424–6439. https://doi.org/10.18653/v1/2021.acl-long.502

[4] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: A Large-Scale Dataset for Educational Question Generation. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). https://doi.org/10.1609/icwsm.v12i1.14987

[5] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[6] Scotty D Craig, Barry Gholson, Matthew Ventura, and Arthur 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 (2000), 242–253.

[7] 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*. Curran Associates Inc., Red Hook, NY, USA.

[8] Huanli Gong, Liangming Pan, and Hengchang Hu. 2022. KHANQ: A Dataset for Generating Deep Questions in Education. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 5925–5938. https://aclanthology.org/2022.coling-1.518

[9] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. https://doi.org/10.18653/v1/P17-1147

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[11] Kalpesh Krishna and Mohit Iyyer. 2019. Generating Question-Answer Hierarchies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2321–2334. https://doi.org/10.18653/v1/P19-1224

[12] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[13] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[14] 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*. Association for Computational Linguistics, Prague, Czech Republic, 228–231. https://aclanthology.org/W07-0734

[15] Seungyeon Lee and Minho Lee. 2022. Type-dependent Prompt CycleQAG : Cycle Consistency for Multi-hop Question Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6301–6314. https://aclanthology.org/2022.coling-1.549

[16] 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742. https://doi.org/10.1162/tacl_a_00343

[19] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2381–2391. https://doi.org/10.18653/v1/D18-1260

[20] Ruslan Mitkov, Halyna Maslak, Tharindu Ranasinghe, Vilelmini Sosoni, et al. 2023. Automatic Generation of Multiple-Choice Test Items from Paragraphs

Using Deep Neural Networks. In *Advancing Natural Language Processing in Educational Assessment*. Routledge, 77–89.

[21] Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence* 12, 1 (2023), 1–32.

[22] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic Graphs for Generating Deep Questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1463–1475. https://doi.org/10.18653/v1/2020.acl-main.135

[23] 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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[24] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, 392–395. https://doi.org/10.18653/v1/W15-3049

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

[26] 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. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[27] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[28] Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics* 44, 3 (09 2018), 393–401. https://doi.org/10.1162/coli_a_00322 arXiv:https://direct.mit.edu/coli/article-pdf/44/3/393/1809172/coli_a_00322.pdf

[29] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics* 8 (2020), 264–280. https://doi.org/10.1162/tacl_a_00313

[30] Luu Anh Tuan, Darsh Shah, and Regina Barzilay. 2020. Capturing Greater Context for Question Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 9065–9072. https://doi.org/10.1609/aaai.v34i05.6440

[31] Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring Question-Specific Rewards for Generating Deep Questions. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 2534–2546. https://doi.org/10.18653/v1/2020.coling-main.228

[32] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[33] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.

[34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[35] 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)*. Association for Computational Linguistics, Hong Kong, China, 6032–6037. https://doi.org/10.18653/v1/D19-1622