From Generation to Evaluation: A Resource for Error-Categorized Question Generation from Video Transcripts

Joshua Berger[1], **Markos Stamatakis**[2], **Anett Hoppe**[2,3,4], **Ralph Ewerth**[2,3,4], **Christian Wartena**[1]
[1]Hochschule Hannover– Data|H Institute for Applied Data Science, Hannover, Germany
[2]TIB– Leibniz Information Centre for Science and Technology, Hannover, Germany
[3]University of Marburg and hessian.AI– Hessian Center for Artifical Intelligence, Marburg, Germany
[4]L3S Research Center– Leibniz University Hannover, Hannover, Germany
{joshua.berger, christian.wartena}@hs-hannover.de
markos.stamatakis@tib.eu, {anett.hoppe, ralph.ewerth}@uni-marburg.de

Abstract

A key challenge in automated question generation is producing grammatically correct, error-free, and contextually relevant questions. While large language models already handle this well, smaller models that can run on consumer-grade hardware face greater difficulties. Another obstacle is the lack of large, high-quality datasets, particularly for education video transcripts, which limits the diversity and applicability of training data. On top of this, current evaluation methods either rely on strict comparison to a "ground truth," undervaluing valid but unmatched questions, or on expert judgments, which do not scale. They do not provide insights into the nature of errors. In this paper, we introduce a dataset of real-life educational video transcripts and investigate the question-generating capabilities of small language models by assessing their output with pre-defined error categories. We also present a novel approach to automatic quality assessment by classifying questions into predefined error categories. We show that questions generated by small language models are still prone to error. Our proposed classification approach outperforms baseline approaches and matches GPT-5 performance by reaching an accuracy of 72%.

Keywords: Educational Video Transcripts, Question Generation, Error Classification

1. Introduction

The rise of online learning platforms and the use of structured assessments have led to a higher frequency of knowledge tests. Many of these tests are designed around question-and-answer formats based on educational video content or lectures, necessitating the manual creation of a substantial number of questions. To mitigate the labor intensity of this process, there is growing interest in automated question generation. State-of-the-art approaches to automatic question generation typically employ large language models (LLMs) such as ChatGPT or Gemini (Anil et al., 2023) to generate questions with a zero-shot approach. Other approaches use fine-tuned small language models (SLMs) like Mistral7B (Jiang et al., 2023) or low-parameter count LLama (Touvron et al., 2023) models to either reduce computation or eliminate the need for API calls. While this is the preferred method from a resource- and cost-efficiency perspective, SLMs only reach a fraction of the performance of their large counterparts. Automatically generated questions face various quality control problems, ranging from dissociation from the source text to more intrinsic problems such as faulty grammar. Additionally, the most common datasets used for training, like SQuAD (Rajpurkar et al., 2016) or NewsQA (Trischler et al., 2017), do not have a focus on education or are taken from exams (e.g., RACE (Lai et al., 2017)) and are not suited to train models for question generation based on videos. Metrics to evaluate questions and identify problems often rely on a comparison to a ground truth (Lin, 2004; Papineni et al., 2002) or are the result of expensive human annotations (Jia et al., 2021; Liu et al., 2020).

In this paper, we investigate typical errors in questions generated by SLMs and address the novel task of identifying various types of issues in generated questions automatically. Our method automates human judgment by training a classifier to recognize different predefined error types. Specifically, we address questions generated from audio transcripts of educational videos. We introduce a new dataset based on transcripts from various English-language educational video content. First, we train multiple small question generation (QG) models, ranging from three to seven billion parameters, on two well-known question-answer datasets, and use them to generate questions for a dataset of educational videos. Subsequently, we manually annotate the questions with error categories. Finally, we train and evaluate several classifiers on the annotated questions.

Thus, we make three main contributions: a dataset of questions of varying quality based on transcripts of educational videos, an in-depth anal-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2166–2177
11–16 May 2026.

ysis of question generation capabilities of state-of-the-art small language models and a classifier to identify common issues in automatic question generation, achieving human-like error analysis. Moreover, the classifier can be used to evaluate generated questions without a ground truth or gold standard. That differs substantially from the prevalent evaluation techniques, which use string matching with a given ground truth. Our classifier runs locally and can be trained and used on a consumer-grade GPU, matching GPT-5 performance while only needing a fraction of the resources. In contrast to most prevalent approaches, we focus on a more specific analysis that has been previously achieved only through human annotation.

2. Related Work

Recent question generation approaches use LLMs to generate questions when given text as input. Question generation using LLMs predominantly uses either fine-tuning (Ko et al., 2020) or zero-shot (Yuan et al., 2023) and few-shot (Poon et al., 2024) techniques to prompt a chatbot model, such as ChatGPT, LLaMA, or Mistral, achieving good results (Elkins et al., 2023; Biancini et al., 2024). However, questions generated with SLMs frequently exhibit errors and do not fully mirror the quality of human-crafted questions.

To assess the degree to which automatically generated questions are correct and useful, several automatic and manual evaluation methods exist (Amidei et al., 2018). The most prevalent automatic metrics are ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), which compare the generated question to a set of reference questions. These metrics are popular due to their ease of use; however, they penalize error-free questions dissimilar to the reference (Nema and Khapra, 2018). To address this limitation, evaluations frequently incorporate manual metrics that use human-generated annotations to assess questions based on criteria such as relevance and grammatical soundness (Mulla and Gharpure, 2023). While manual metrics offer superior assessments of question quality, they also require more time and labor.

The primary approach to address these limitations involves methods that automate manual metrics and endeavor to emulate human judgments. Examples include *Learned Evaluation metric for Reading Comprehension* (LERC) (Chen et al., 2020) and *Reference-free Question Generation Evaluation* (RQUGE) (Mohammadshahi et al., 2023). LERC, a BERT-base model, uses a scoring mechanism to evaluate how well a question is suited to test text comprehension. RQUGE, on the other hand, uses a combination of question-answering and span-scoring modules to determine the answerability of questions.

Other metrics include *Crosslingual Optimized Metric for Evaluation of Translation* (COMET) (Rei et al., 2020), BLEURT (Sellam et al., 2020) and *Compression, Transduction, and Creation* (CTC) (Deng et al., 2021). These metrics automate human judgments on a wide variety of text generation tasks, like machine translation.

While these metrics are mostly regression-based and assign a singular score to a candidate question, our approach takes a different angle. We approach it as a classification task to identify specific problems in the candidate, providing a detailed analysis by isolating potential problems.

Datasets used for question generation are mostly based on text. The most popular datasets, SQuAD (Rajpurkar et al., 2016), RACE (Lai et al., 2017) and NewsQA (Trischler et al., 2017), all use written text in the form of Wikipedia entries or news articles as a basis. While this often makes sense, it overlooks a significant aspect of modern education: video content. Existing video transcript datasets often either do not feature a question component (e.g. HD-VILA-100M (Xue et al., 2022), HowTo100M (Miech et al., 2019) or MSR-VTT (Xu et al., 2016)) or do not have an educational focus (e.g. CinePile (Colas et al., 2020a) or MovieQA (Tapaswi et al., 2016)). The few video transcript datasets that exist, such as TutorialVQA (Colas et al., 2020b), are primarily based on short-form instructional videos. This is where our dataset TIB-AV-100 fills a gap by being a video transcript dataset that provides questions for long-form and complex educational videos, ranging from typical lecture and whiteboard scenarios all the way to documentaries.

3. Datasets

In this section, we describe the datasets used. This includes the dataset we introduce, TIB-AV-100, as well as previously available datasets used for training and evaluation.

3.1. TIB-AV-100: A Video Transcript Dataset

To obtain a dataset with a large variation in questions, question types, and types of errors, we trained various state-of-the-art small language models to generate questions: a Gemma 3 (Kamath et al., 2025) model (4B parameters), a LLama 3.2 (Dubey et al., 2024) model (3B parameters) and a Deepseek (DeepSeek-AI et al., 2025) model (7B parameters), distilled on Qwen. All of these models were fine-tuned on the SQuAD dataset with the recommended 80/10/10 standard split. The models were mostly finetuned with default set-

tings. The only customized ones were a per-device train and eval batch size of two and a runtime of only three epochs. To fit the models on a single GPU, we used LoRA for causal LMs with standard settings. To avoid monotonous question generation and diversify our dataset, especially regarding question complexity and wording, we fine-tuned another set of these models on the more complex NewsQA (using the same model settings as SQuAD). Also, we used the models' instruction-fine-tuned variants for zero-shot question generation. Counting zero-shot models, this results in a total of nine different question generation methods. This mix of models and training datasets allowed us to obtain a wide variety of questions. To get the best possible zero-shot results, we prompted the models using best practices (Chen et al., 2025) and prompt guides [1,2] with the following prompt:

```
Consider this audio transcript:
{CONTEXT HERE}

You are a professor creating
questions for an exam geared at
college students.

Generate a single question that
tests the students understanding
while also not demanding an
overly long or complex answer.
Make sure the question can be
answered with the context.  Do
not generate an answer and
generate nothing but the question
itself.
```

This prompt was found through experimentation over several iterations and was found to yield the best results. We then used these models to generate questions for 4,700 text chunks from the TIB-AV-100 transcripts, resulting in about 42,300 questions with nine questions per text chunk. The dataset can be found here.

All training was performed using an Nvidia RTX3090 24GB. Overall, training plus final question generation took about 72 GPU hours per model, resulting in a total cost of 432 GPU hours.

3,000 generated questions were then manually annotated by a native English speaker for typical error types and used to train and evaluate the error classification algorithms. A more detailed description of the annotation process and analysis of annotation results, as well as dataset composition and examples, can be found in Section 4.

3.2. Error Classification

Our error classification pipeline utilized the following datasets:

Stanford Question Answer Dataset (SQuAD) (Rajpurkar et al., 2016) consists of about 100,000 English triples (text, question, answer) from Wikipedia. The SQuAD dataset is popular in question-related text generation tasks, and we used it to train our models.

News Question Answering (NewsQA) (Trischler et al., 2017) consists of over 100,000 English question-answer pairs annotated from over 10,000 news articles from CNN. We used this dataset as a second basis of training data to train a second set of models.

TIB-AV-100 consists of the top 100 English videos of the video archive of the *Leibniz Information Centre for Science and Technology* (TIB)[3]. We transcribed these videos using OpenAI's Whisper (Radford et al., 2023) (*large-v3*) and then segmented the transcripts into chunks of about 160 words, approximately the chunk size of the SQuAD texts. Cut-off sentences were restored, leading to slight overlaps in some chunks. TIB-AV-100 was used as a basis to train the classifier and evaluate its performance. This dataset was created to closely mirror actual real-world tasks for which automatic question generation is used.

Fairytale Question Answering (FairytaleQA (Xu et al., 2022)) consists of 10,580 English questions about 278 famous fairytales. This dataset aims to test the narrative comprehension of children and consists of relatively easy-to-understand texts. We used it as a second evaluation dataset for the classifier to test knowledge transfer performance.

4. Error Detection

In this section, we first outline the relevant error categories for the classifier and then show how they are distributed in the dataset.

4.1. Error Categories

We identified ten typical problems with automatically generated questions by manually analyzing a subset of the generated data. These ten specific issues can be grouped into five broader categories as shown in Table 1 (category 1 - 5). We extended these ten categories to also handle questions that fall outside their scope by adding two more: *Other error*, to include labels for questions that contain errors not covered by the ten categories, and *Error-free*, to label questions that do not contain errors.

We manually evaluated 2,987 questions generated by the mentioned models using text chunks from the TIB-AV-100 dataset and annotated questions with these twelve categories. In addition to labeling the questions, we also annotated the text

[1] https://www.promptingguide.ai

[2] https://platform.openai.com/docs/guides/text

[3] https://av.tib.eu/

Cat.	Description and Sub-Categories	Example
0	Error-free	What is the name of the renderer that produces vector tiles?
1	Does not fit the text (a) Can't be answered due to missing information in the text (b) Irrelevant or off-topic	How do I get sharper images? *for a text about cutting wedges*
2	Can't be understood without reference to the text	How do I integrate this in three seconds? (better: *How do I integrate the rectangle surface in three seconds?*)
3	Nonsensical	How do I make love not poor not TV?
4	Asks for more than one thing or is vague	What happens when you get into a collision? (better: *What is the most common injury in a collision accident?*)
5	Unnatural or wrong phrasing (a) Grammatically incorrect (b) Unnecessarily long (c) Wrong pronoun (d) Wrong question word (e) Unusual wording	What is the square root zero? (better: *What is the square root of zero?*)
6	Other error	How about asking about the impact of the alleles on the colony? (better: *What is the impact of alleles on the colony?*)

Table 1: Error categories used for annotation and their sub-categories. Examples were chosen from questions generated for the TIB-AV-100 data.

chunks from the TIB-AV-100 dataset as either *good* (i.e. suitable to ask a question about) or *bad* (i.e. not worthy to ask a question about). An example of what differentiates a good from a bad context can be seen in Table 2. Questions were then annotated only for the text chunks annotated as good. Except for category 0, all categories can be part of a multi-label set per question. Additionally, our annotator was tasked with manually creating an error-free question for each of the annotated good text chunks. A native English speaker with a background in linguistics annotated all questions.

4.2. Error Distribution

Figure 1 shows the distribution of error types in the training and test data. The data exhibits significant error variations, resulting in a diverse set of generated questions. In total, 2,181 error labels were assigned to 1,577 questions. 47.2% of questions are error-free, with about a quarter of them being the manually created ground truth questions. The most common errors are: questions that can't be answered (22%), questions that are vague (14.9%) and questions that are worded unusually (8.8%). Additionally, most questions were only labeled with one error type, as seen in Figure 2. Further examination of the data shows that there are also error differences between models and methods, shown in Figure 3 and Figure 4. Deepseek created the most error-prone questions out of all models. While Gemma created the most answerable and error-free questions, this model also created the vaguest questions. Regarding the methods, SQuAD seems to be the most reliable training source, while NewsQA created the least error-free questions. Zero-shot shows good capabilities at creating answerable questions, but also suffers from the problem of creating the vaguest questions by far. All methods and models created almost no questions that are nonsensical or can't be understood without reference to the text. Additionally, models and methods did not create many questions that fall outside of the examined error categories (only 6.1% of questions contained other errors), showing that the used error categories cover almost all errors present in questions generated by SLMs.

5. Error Classification

In this section, we discuss the classifiers we used for error classification and the training process.

Good	Bad
Research is a key pillar of social progress. Worldwide, over two trillion euros are invested in acquiring new scholarly knowledge annually. Yet we have used roughly the same method of scholarly knowledge communication since the birth of modern science. Scientific articles. A 17th century researcher would have been able to read the whole scientific canon. [...]	[...] I will use some time on GitHub, just looking to the code, showing everywhere, everyone where stuff is. And also, hopefully, I will be able to do a couple of live demos. Knock on wood. So, yeah, I'm going to try to summarize afterwards by going through good, bad, ugly, and also try to sketch out the future, as I see it. And end up with a summary. [...]

Table 2: Examples of good and bad texts from the TIB-AV-100 dataset. These examples were chosen to be representative of the texts encountered.

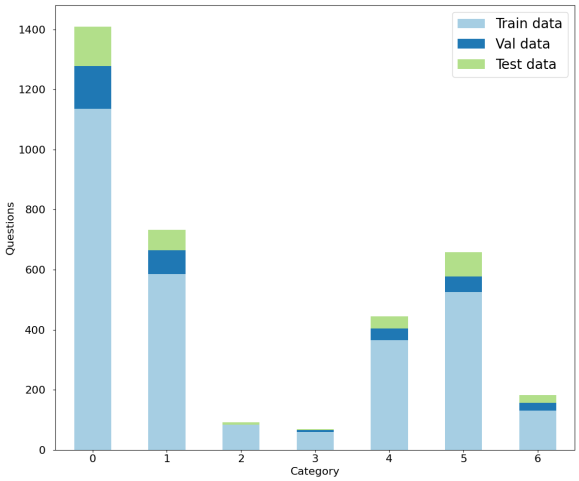

Figure 1: Number of questions per error category. 0 = error-free; 1 - 6 = error categories.

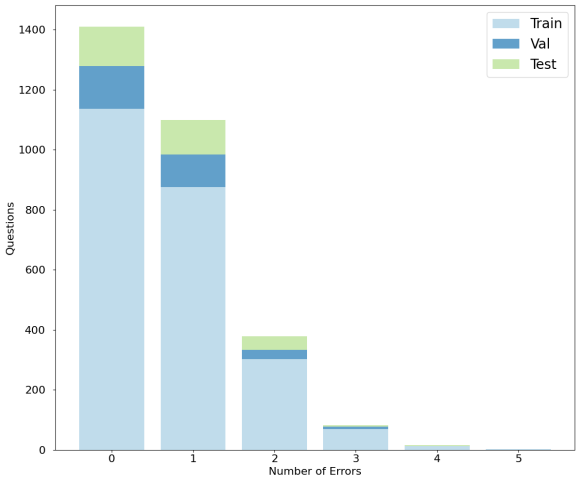

Figure 2: Number of questions per error amount.

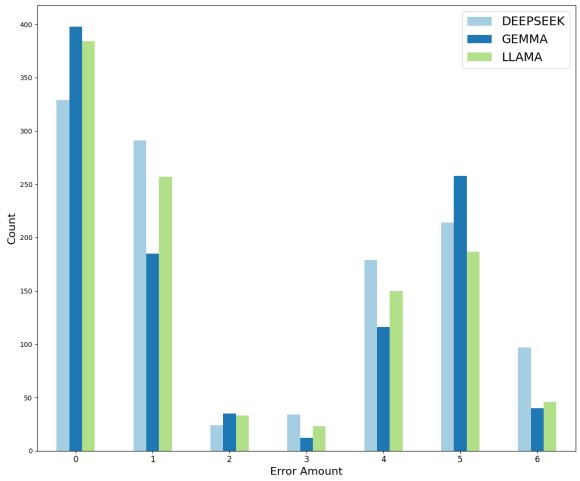

Figure 3: Total # of errors per model and error amount.

5.1. Classifier Training

We consider two classification tasks: a binary classification to identify error-free questions and a multi-label classification that assigns questions to five broader error categories, as well as the "Other error" and "Error-free" categories. For both problems, we trained multiple state-of-the-art BERT-based encoders: BERT (Devlin et al., 2019) (*bert-large-cased*), DeBERTa (He et al., 2021) (*deberta-v3-base*), XLNet (Yang et al., 2019)(*xlnet-base-cased*), NomicBERT (Nussbaum et al., 2024) (*nomic-bert-2048*) and GTE (Li et al., 2023) (*gte-large*). To perform classification, we extended these encoders by adding a classification layer on top that takes the pooler output from the encoder and projects it through a linear layer onto the dimensionality of the number of output labels. In this case, two labels are used for binary classification, and seven labels are used for multi-label classification. To find the optimal training strategy, we experimented with batch size, learning rate and number of epochs. All models took a combination of text and the corresponding question as input. Since the text chunks are rather short, truncating the input was not necessary. Instead, it was padded to the context window of the specific models. Where applicable, text and question were separated by a SEP token. We trained classifiers for both tasks separately. Due to the increase in complexity for the multi-label task, as well as the different natures of the classifiers regarding pre-training and archi-

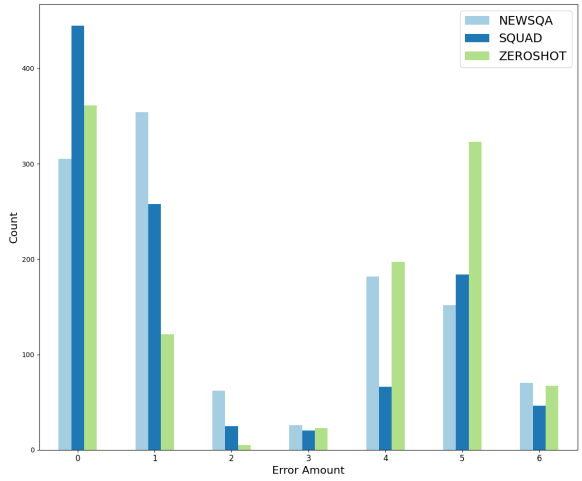

Figure 4: Total # of errors per method and error amount. SQUAD and NEWSQA refer to the fine-tuned models, while ZEROSHOT refers to the prompted models.

tecture, training separately for binary classification might give a performance boost. Additionally, all classifiers leave a small enough footprint on the hard drive that storing two of them is negligible. The 2,987 annotated questions are split randomly into a training set of 2,397, a validation set of 294 and a test set of 297 questions. Since we generated multiple questions per text chunk, we split the dataset context-wise. Thus, ensuring all questions for a text chunk remain in the same split. Overall, we arrived at a learning rate of $2 \cdot 10^{-6}$ over ten epochs. We applied a standard weight decay of 0.01 and left dropout on hidden and attention layers at the standard value of 0.01. The models were trained with a batch size of 32, both for training and evaluation. All models were trained using cross-entropy loss on the binary classification task. For the multi-label classification task, we used binary classifiers, one for each label, which were trained using a binary cross-entropy loss function with class weights. For simplicity's sake, these multiple binary classifiers will be referred to as *multilabel classifier* going forward.

All classifier training was performed using an Nvidia RTX3090 24GB. Due to the small amount of data, it took < 1 GPU hour to train the final models. All classifiers were trained on GPUs, with overall experimentation costing about 48 GPU hours.

5.2. Zero-Shot GPT-5

In addition to training classifiers, we also performed a zero-shot classification with GPT-5. This is meant to serve as a baseline performance that can be achieved even without any training data and by just using the metric, a text and questions by themselves. We iterated over multiple prompts but finally landed on the following prompt:

```
You are a professor labeling
questions based on their
usability in a college exam.

The questions should be labeled
regarding the following error
categories:
{LIST OF CATEGORIES 1 - 12 HERE}

Each question should be sorted
into one or more of these
categories.  Category 12 is
exclusive.  Questions in that
category can not be in other
categories.

For the following context:
{CONTEXT HERE}

Label the following questions:
{QUESTION LIST HERE}

Format your answer like this:
Question:  [Labels]

Generate only the answer in the
given format and nothing else.
Do not justify or explain your
answer.
```

This prompt is similar to the instructions given to the annotators who labeled the dataset. The list of categories used in this prompt is the one presented in Table 1, but treats the sub-categories as their own category, resulting in a total of twelve individual categories. The question list consists of all questions generated for a context. We used the same test dataset for the GPT-5 experiment that we also used for the other experiments. Each question list was provided to the model as part of a new prompt.

6. Classifier Evaluation

In this section, we evaluate the performance of the trained classifiers separated into classification approaches. Table 3 shows the results for the binary classification. We can see that zero-shot GPT-5 already performs reasonably well, reaching an accuracy of 72%. Our fine-tuned classifiers differ greatly in performance. While the worst one, GTE, even underperforms a simple majority classifier, NomicBERT manages to match the GPT model, while XLNet almost reaches its performance. Interestingly enough, all fine-tuned models perform slightly better on the minority class than the majority class, except for NomicBERT. Regarding the weighted F1 score, all classifiers vastly outperform the majority classifier. XLNet nearly reaches the zero-shot performance of the GPT-5 model, while NomicBert matches it.

In addition to evaluating on a test split of the TIB-AV dataset, we also tested on 222 manually

Clf.	Class	Prec.	Rec.	F1	F1	Acc.
BERT	0	0.58	0.67	0.62	0.64	0.64
	1	0.70	0.61	0.66		
DB	0	0.76	0.29	0.42	0.60	0.65
	1	0.62	0.93	0.75		
GTE	0	0.44	0.47	0.47	0.50	0.50
	1	0.56	0.52	0.54		
NB	0	0.63	0.89	0.74	**0.72**	**0.72**
	1	0.87	0.58	0.70		
XLNet	0	0.67	0.64	0.65	0.69	0.70
	1	0.73	0.75	0.74		
Maj.	0	0.53	1.00	0.69	0.35	0.53
	1	0	0	0		
GPT	0	0.68	0.79	0.73	0.72	0.72
	1	0.76	0.65	0.70		

Table 3: Results for the best binary classifiers, including a baseline majority classifier. NB = NomicBERT; DB = DeBERTa. 0 = error-free; 1 = error. The second F1 value is the weighted average for each classifier. Results are for the TIB-AV dataset.

Class	Prec.	Rec.	F1	Support
0	0.58	0.61	0.60	132
1	0.29	0.07	0.11	61
2	0.00	0.00	0.00	8
3	0.00	0.00	0.00	3
4	0.44	0.51	0.47	41
5	0.81	0.35	0.49	71
6	0.00	0.00	0.00	27
Weighted			0.41	

Table 4: Results for the multilabel classifier (*bert-large-cased*). 0 = error-free; 1 - 5 = error categories. 6 = Other error. Results are for the TIB-AV dataset.

annotated questions generated on the FairytaleQA dataset. Transferring knowledge to a different domain remains a challenging task for the classifiers, as can be seen in Table 5. While all classifiers except for BERT manage to outperform the majority baseline, only XLNet has noticeable gains over the baseline. BERT and DeBERTa both classified most encountered questions as errors, showing that they did not learn meaningful representations. While XLNet still tended to favor the error class, NomicBert is the only classifier to favor the minority class (error-free), indicating learned behaviour from its training dataset. When comparing class F1 scores with the class F1 scores in Table 3, we can see that all models stick to their preference in label.

The multilabel results show that multilabel classification is a much more complex task than binary classification. This is evident even in the GPT-5 baseline model. This model does not achieve compelling results on the task, reaching only a weighted F1 score of 0.55, as can be seen in Table 6. Here, we can see that GPT-5 performs well on the error-free class but becomes significantly worse when tasked with identifying specific issues with the questions. Especially, error categories 2 (*can't be understood without reference to the text*) and 3 (*nonsensical*) have not been identified by the model a single time, possibly due to the low amount of support examples. For actual error detection, GPT-5 performs best on error category 1 (*does not fit the text*).

The results for the best fine-tuned multilabel classifier (*bert-large-cased*) can be seen in Table 4. When compared to the GPT-5 baseline, we can see that the fine-tuned classifier also never predicts categories 2 and 3. The fine-tuned classifier

Clf.	Class	Prec.	Rec.	F1	F1	Acc.
BERT	0	0.12	0.03	0.05	0.35	0.45
	1	0.49	0.82	0.62		
DB	0	0.75	0.05	0.06	0.40	0.55
	1	0.54	0.99	0.70		
GTE	0	0.56	0.50	0.53	0.58	0.59
	1	0.60	0.66	0.63		
NB	0	0.54	0.75	0.63	0.58	0.59
	1	0.67	0.45	0.54		
XLNet	0	0.73	0.37	0.49	**0.62**	**0.64**
	1	0.62	0.88	0.73		
Maj.	0	0	0	0	0.37	0.54
	1	0.54	1.00	0.70		

Table 5: Results for the best binary classifiers, including a baseline majority classifier. NB = NomicBERT; DB = DeBERTa. 0 = error-free; 1 = error. The second F1 value is the weighted average for each classifier. Results are for the FairytaleQA dataset.

Class	Prec.	Rec.	F1	Support
0	0.68	0.78	0.73	132
1	0.56	0.63	0.59	61
2	0.00	0.00	0.00	8
3	0.00	0.00	0.00	3
4	0.72	0.47	0.57	41
5	0.41	0.27	0.32	71
6	1.00	0.08	0.15	27
Weighted			0.55	

Table 6: Results for the GPT-5 multilabel classification. 0 = error-free; 1 - 5 = error categories. 6 = Other error.

additionally never predicts category 6 (*other error*). Table 1 shows that these three categories are by far the categories that have the least amount of support. On the other categories, the multilabel classifier does not reach the performance of GPT-5, except for category 5 (*unnatural phrasing*), where it outperforms GPT-5 by a significant margin. Overall, results show the validity of treating binary and multilabel classification separately, since each task achieved best results with another classifier.

7. Discussion and Conclusion

In this paper, we introduce a dataset of comprehension questions derived from transcripts of educational videos, created using various question generation models. The study focuses on analyzing typical errors found in questions produced by state-of-the-art small language models. A total of 2,987 questions were manually annotated for ten common error types, alongside a broader "Other errors" category and a label for error-free questions. Furthermore, a classifier is proposed to automatically detect specific types of errors in generated questions.

We trained five classifiers on the binary task to predict whether a question is error-free and on the multilabel task of finding the exact types of errors for each question.

The binary classification results show that the classifiers can detect the presence of errors in automatically generated questions to a degree that enables their use in downstream tasks. Such tasks could be a reward model in reinforcement learning or a final filtering step to extract high-quality questions from a pool of candidates. The classifier could be used as part of an evaluation strategy for question-generation models. The performance of the classifiers approaches that of a zero-shot GPT-5 approach, while one of them matches GPT-5. This is achieved on a local machine using only a fraction of the compute cost of GPT-5. Our classifier also does not require a ground truth question to compare a generated one against. This offers an advantage over most commonly used automatic evaluation methods, which rely on comparison to a ground truth and fail to detect subtle grammatical errors while unfairly penalizing valid questions that differ from the ground truth phrasing.

The multi-class multi-label task of detecting the exact set of problems is much harder and the results are not yet compelling. However, this fault could be attributed to the low amount of available training data and not the method or classifier itself. Furthermore, GPT-5 also does not achieve compelling results in the multilabel classification task, indicating the complexity of the task.

Future work includes expanding datasets, improving classifier performance, and integrating this method into an automated assessment pipeline.

Acknowledgements

This work was funded by the Lower Saxony Ministry of Science and Culture with funds from the zukunft.niedersachsen program of the Volkswagen-Stiftung as part of the VidQA project.

Ethical Considerations

We ensure that every step taken in the creation of this paper was subject to ethical scrutiny. All models and datasets used in this paper were either self-created or used as intended and under their licenses. Every mention of related research and data is properly cited. The annotator who provided work for this paper was properly employed and paid for by our institution. The annotator was aware of the nature of their task and agreed to the use of their data.

The work presented in this paper is our own and AI assistance was only used to help with formatting and formulating the text.

Limitations

While the binary classifier shows promise, several limitations remain to be considered. The small number of training samples led to a severe underrepresentation of several classes. The classifier itself can not distinguish errors not identified in this paper, as they were caught under the umbrella category "Other errors".

8. Bibliographical References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 307–317. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald

Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. 2024. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, page 584–590, New York, NY, USA. Association for Computing Machinery.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6521–6532. Association for Computational Linguistics.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, page 101260.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2020a. Tutorialvqa: Question answering dataset for tutorial videos.

Anthony M. Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2020b. Tutorialvqa: Question answering dataset for tutorial videos. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5450–5455.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7580–7605. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank

Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie Chi Kit Cheung. 2023. How useful are educational questions generated by large language models? In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky - 24th International Conference, AIED 2023, Tokyo, Japan, July 3-7, 2023, Proceedings*, volume 1831 of *Communications in Computer and Information Science*, pages 536–542. Springer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2021. EQG-RACE: examination-type question generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13143–13151.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, and Ivan Nardini. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6544–6555. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Chin-Yew Lin. 2004. ROUGE: A package for auto-

matic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhuang Liu, Kaiyu Huang, Degen Huang, and Jun Zhao. 2020. Semantics-reinforced networks for question generation. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, pages 2078–2084.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6845–6867. Association for Computational Linguistics.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Prog. in Artif. Intell.*, 12(1):1–32.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3950–3959. Association for Computational Linguistics.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *CoRR*, abs/2402.01613.

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

Yin Poon, John Sie Yuen Lee, Yu Yan Lam, Wing Lam Suen, Elsie Li Chen Ong, and Samuel Kai Wah Chu. 2024. Few-shot question generation for reading comprehension. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 21–27, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5026–5035. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained llms: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12952–12965. Association for Computational Linguistics.