EduQuest: Lecture Texts and Questions for Higher Education

Anonymous ACL submission

Abstract

Question generation (QG) techniques carry great educational potential for producing various learning materials and measuring student understanding. However, existing datasets for building QG approaches predominantly feature simpler texts and exercises aimed at a younger audience, which engage little higher-order thinking, thereby limiting their suitability for developing question-generation tools tailored to higher education. Additionally, they often originate from only one or two sources, resulting in low diversity and variety. We introduce EDUQUEST, which directly addresses these limitations by integrating a collection of open-source textbooks, lesson notes, tests, and exercises for higher education from OpenStax and OpenText, MIT OCW, CK12, and KhanAcademy, combining diverse learning materials and teaching methodologies from various disciplines and educators. Moreover, the dataset provides various meta-features, such as question types and difficulty levels, allowing customized question generation to accommodate instructor needs. Experimental results prove that models trained on EDUQUEST can generate high-quality and educationally useful questions relevant to the material.

1 Introduction

Education research has demonstrated that active learning is the most efficient method of learning (Prince, 2004). In this context, high-quality questions play a crucial role in enabling learners to gauge their understanding of subjects and trigger critical thinking accurately.

Creating well-designed questions can be a demanding and time-intensive task. Educators must employ various question types with different difficulty levels to tailor questions and exercises to their students' needs. Additionally, questions and exercises should be clear and logically structured to enable students to focus on the task at hand while being distinctive and challenging enough to prompt them to think critically about what they know and how to apply it, which is especially important in higher education.

To address this challenge, we propose a comprehensive novel dataset and provide a tool built with this dataset capable of generating complex questions from educational texts.

The dataset addresses the following key limitations in existing datasets for educational question generation and summarization:

- Limited Diversity/Topics;
- Not specifically designed for educational purposes;
- Simple questions that do not require higher-order cognitive skills to solve and only require lookups on the text.

The issue of limited diversity is prevalent. Most question-answering datasets, such as SQuAD (Rajpurkar et al., 2016), are collected solely from a few sources, such as Wikipedia, resulting in a constrained topic range and a homogeneous writing style. Similarly, many existing educational text datasets lack diversity as they, too, originate only from a single or a handful of sources, restricting the breadth of topics and styles.

The third limitation concerns the Purpose of Design. SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017), for instance, were not explicitly crafted with an educational aim in mind, leading to limited usefulness in the educational context due to the inherent differences to educational texts, exercises, and questions.

Lastly, most existing datasets only contain simple questions, which do not require higher-order cognitive skills. These questions generally only require the student to remember or look up the answer in the text, failing to emulate the complexity and challenge of questions that engage the user to

think about the subject, typically found in higher educational exams and textbooks.

We thus present *EduQuest*, a comprehensive novel dataset that expands the scope of educational question generation datasets to encompass higher educational texts. By carefully gathering lecture-question pairs authored by domain experts from OpenStax, OpenText, MIT OCW, CK12, and KhanAcademy, *EduQuest* offers a collection of 76008 lesson documents, along with 68248 corresponding questions and exercises. This extensive compilation covers various subjects, such as STEM, social sciences, and more, while incorporating elementary questions, targeting basic reading comprehension, and complex questions requiring higher-order cognitive skills and idea association. In addition, *EduQuest* also provides the difficulty and question type classification in the revised Bloom's Taxonomy (Bloom and Krathwohl, 2020).

We trained and evaluated state-of-the-art networks with proven performances on question generation and summarization tasks to demonstrate the *EduQuest's* effectiveness in training deep learning models to perform the QG task on higher-learning lesson texts. The results indicate that the QG models learned to generate diverse, high-quality questions and exercises from complex higher educational texts.

2 Related Work

2.1 General Datasets for Question Generation

Squad (Rajpurkar et al., 2016) is a dataset composed of questions generated by online crowdworkers and can always be answered by simple lookups on the accompanying text. Despite its breadth of topics, the questions may not be the quality one would expect from a teacher or lecturer, belonging primarily to the lowest levels of the revised Bloom's Taxonomy, Remembering, and factual knowledge. Similarly, **TriviaQA** (Joshi et al., 2017) is composed of questions taken from online trivia websites, suitable for accessing knowledge over a wide range of subjects but not usable for educational purposes.

2.2 Educational Datasets for Question Generation

The **Textbook Question Answering** (TQA) dataset (Kembhavi et al., 2017), launched by the Allen Institute for AI in 2017, is an extensive dataset tailored for research in Multi-Modal Machine Comprehension (M3C). This dataset, while comprehensive and of high quality, is primarily derived from middle school science curricula. While the TQA dataset is a valuable resource for middle school-level education, its utility for higher education is limited. The questions' simplicity and the lessons' elementary nature make them less applicable to advanced educational settings. Similarly, **ScienceQA** (Lu et al., 2022) suffers from similar issues.

LearningQ (Chen et al., 2018) is a popular educational question generation dataset, built from data from Khan Academy[1] and TED-Ed[2]. Their variability is limited despite covering a broad spectrum of subjects since their questions come from only two sources. Furthermore, while it contains high-quality questions designed by educational experts from TED-Ed, these questions' corresponding lecture texts are transcripts from videos which are often different in nature from actual educational texts. The vast majority of questions are a collection of audience's comments on the videos and articles that included a question mark. Many of which are not directly relevant to the corresponding lesson. Moreover, in its current state, and unlike *EduQuest*, it is challenging to use LearningQ in a plug-and-play fashion due to the substantial amount of noise in the form of unprocessed texts filled with escape characters and markdown syntax or emoji codes.

FairyTaleQA (Xu et al., 2022) consists of over 10k explicit and implicit question-and-answer pairs associated with children's stories. The quality of the questions in this dataset is very high because education experts crafted them, and the dataset is a valuable addition to the field. However, because the questions were specifically designed for young readers, they are primarily composed of easy-to-grasp language and words, and the texts and questions lack the complexity found in texts and questions for higher education levels.

EduQuest addresses these issues by combining educational texts designed by education experts for higher education topics from different sources, ensuring quality and fidelity.

3 The EDUQUEST Dataset

In this section, a comprehensive exposition of the dataset's structure, including its collection methodology, is provided. Subsequently, an analysis of the said dataset is conducted and presented.

[1] https://www.khanacademy.org/
[2] https://ed.ted.com/

	Lessons/Books	Questions	Highest_level
EduQuest	**76008**	68248	**graduate-school**
LearningQ	10841	**231470**	high-school
FairytaleQA	278	10580	elementary-school
SixthGrader	1076	26260	elementary-school

Table 1: Comparison of EduQuest with related datasets, by number of lectures, questions, and by how advanced the materials are. Highest_level indicates the most advanced educational level lessons present in the dataset

	Books	Lessons	Exercises	Token Length Mean	Token Length Std
Openstax	36	3723	15787	461.5	1422.6
Opentext	41	1072	910	1187.5	4218.2
CK12	14	2164	6695	188.5	365.6
MIT OCW	338	338	5959	68.3	246.9
Khan Acad.	3031	102819	1873	70.6	124.0

Table 2: Overview of the Dataset

3.1 Source Texts

EduQuest drew upon five diverse and resource-rich repositories: OpenStax, OpenText, MIT OpenCourseWare (OCW), Khan Academy, and CK-12. These data sources were selected for their comprehensive coverage across various academic disciplines and commitment to open-access education. This unique blend of resources not only enhances the robustness of our dataset but also caters to diverse learning styles and educational needs. Each data source has unique properties that influence the educational content's type, format, and style. The characteristics of these sources provide our dataset with a rich and diverse range of educational texts, questions, and exercises. The following sections will detail the properties of each data source, their respective contribution to our dataset, and how these properties inform the overall design and potential uses *EduQuest*.

3.1.1 OpenStax

OpenStax[3] is a nonprofit educational initiative based at Rice University that publishes high-quality, peer-reviewed, openly licensed textbooks for college and high school courses. Their textbooks can be used at no cost and are backed by additional learning resources. OpenStax textbooks cover a wide range of subjects, including STEM, social sciences, and others. Textbooks were filtered based on their complexity and suitability for high school level and up. Simultaneously we selected texts rich in textual context, and we filtered out subjects of which the majority of problems were equation-based with little or no textual context, such as Calculus and Algebra. The selected texts were then downloaded from the OpenStax Website. Due to the excellent quality of the textbooks, which were meticulously designed for both self-study and instructional use, they are richly structured with clear formatting, which allowed us to mark many of the question types—in addition to the presence of questions at the end of sections, chapters, and spanning the entire textbook results in a versatile and diverse dataset of lesson and question pairs for modular and specific use cases.

3.1.2 OpenTextBC

OpenTextBC[4] is a project of the British Columbia Ministry of Advanced Education, Skills, and Training that provides free, open-source textbooks for post-secondary courses, with textbooks covering a variety of subjects, including STEM, practical skills, and others. We selected and collected a subset of textbooks from the extensive OpentTextBC Catalog, including high-quality questions and learning objectives that could easily be extracted and processed. Many of these textbooks also included learning objectives which were marked and extracted. The practical skill textbooks add valuable subject and style variety to *EduQuest* lesson texts.

[3]https://openstax.org/

[4]https://opentextbc.ca/

3.1.3 KhanAcademy

Khan Academy[5] is a non-profit educational organization that offers free, personalized learning resources for all ages, covering math, science, computer programming, history, art history, and economics. The platform provides practice exercises, instructional videos, and a personalized learning dashboard that allows learners to study at their own pace, both in and outside the classroom. The Khan Academy lessons and questions were sourced from the LearningQ dataset. The dataset initially did not include instructor-posed questions from the Khan Academy platform but only Lessons and comments from users that included a question mark. Upon careful analysis, however, we found that many of the lessons were noisy, containing embedded instructor questions that could be processed and extracted from the lesson narratives. This meticulous process resulted in higher-quality questions and lesson pairs with clear separation. The post-processing also included removing artifacts and useless questions. Because we found the learner comments often not relevant or of high quality, we should have included or processed them. Orphaned lessons without associated questions were flagged as such but not deleted. Lessons and texts for high school and above were marked, enabling flexible selection for additional research and use purposes.

3.1.4 MIT OpenCourseware (OCW)

MIT OCW[6] is a free, publicly accessible, and openly-licensed digital collection of high-quality teaching and learning materials from the Massachusetts Institute of Technology. It covers virtually all MIT course content and is a permanent MIT activity. OCW offers course materials from thousands of MIT courses, covering the entire MIT curriculum, and does not offer credit or certification to users The materials include lecture videos, written assignments, lecture notes, problem sets with solutions, and exams with solutions.

After scraping the contents of MIT OCW, we obtain the unprocessed text corresponding to 328 lectures, that contain either assignments, exams, or both. Afterward, we post-process the acquired assignments and exams to find the questions using GPT-3.5, obtaining 5959 questions. Given their nature, each scraped question from OCW is significantly more involved than the ones previously obtained, often having several interlinked subquestions (Table 3) within it that require an involved thought process to solve. The exam questions are self-contained with respect to the lecture material, whereas the assignment questions might be more challenging and involved.

3.1.5 CK12

CK-12[7] is a non-profit organisation dedicated to increasing access to high-quality educational materials for K-12 students worldwide. It offers free, standards-aligned, open content in STEM subjects (Science, Technology, Engineering, and Math). CK-12 provides a variety of resources, including interactive practice problems, articles, videos, and teacher-directed classes for whole classes or smaller, differentiated groups to personalise learning.

3.2 Question Annotation

EduQuest includes meta-features to allow for more variance in the question-generation process. It has been shown (Du et al., 2020) that these features can improve the performance of models in question generation and question-answering, thus motivating this decision.

In order to facilitate generating specific types of questions, thus providing more flexibility in the question generation process, most questions in *EduQuest* are labeled with their respective question type. Definitions of question types are shown below, with examples presented in the Appendix.

1. **Multiple Choice** questions present test takers with a problem and a set of possible answers, with only one being correct. The task is thus to find the correct statement amongst the wrong ones.

2. **True or False** questions consist of one statement and ask test takers whether that statement is right or wrong regarding the source lecture.

3. **Fill the Blank** questions present the test taker with an incomplete sentence and ask the user to complete it with information present in the source lecture. These types of questions have an intersection with multiple-choice questions.

[5] https://www.khanacademy.org/
[6] https://ocw.mit.edu/
[7] https://www.ck12.org/student/

Question 2 [60 points] (a)If the aggregate technology exhibits constant returns with respect to the vector of accumulable factors (different types of capital), then the economy has necessarily a constant growth rate at all times, and it is impossible to make sense of conditional convergence. [15 points] (b) More competition necessarily promotes economic growth and social welfare, since firms are forced to produce more goods and extract less pro fits from consumers. [15 points] (c)Consider an individual agent. If her income varies randomly from one period to another, then her consumption will also vary from one period to another, but less so than her income. [15 points] (d) The neoclassical growth model (the RBC paradigm) can well account for the business-cycle variation in output, investment, employment, and total factor productivity. [15 points]

Table 3: Example of an OCW question with several parts, corresponding to the course 14-05, Intermediate Macroeconomics, Spring 2013.

4. **Concept** questions are straightforward, usually requiring the test taker to recall a definition or phrase in the source document.

5. **Open Ended** questions typically require a longer answer than the other four question types. Open-ended questions are those that allow someone to give a free-form answer, requiring students to either reexamine text evidence or extend their own thinking,

The labeling was conducted manually, either by the authors of the source text or afterward during the dataset processing. These labels allow for generating various question types that target different skills so that the models trained on this dataset can also increase their variety.

Furthermore, every question present in the dataset also has been classified in the **cognitive process** and **knowledge** dimensions of the revised Bloom's taxonomy (Bloom and Krathwohl, 2020; Krathwohl, 2002), indicating the expected learning objectives of each question among two dimensions. In the cognitive process dimension, each question is classified into the categories increasing in cognitive complexity described below.

- **Remember** questions require the test taker to retrieve relevant knowledge from long-term memory - usually a direct concept or definition;

- **Understand** questions ask the user to construct meaning from some source text;

- **Apply** problems ask the test taker to apply some method directly explained in the source text;

- **Analyze** questions ask to break material into foundational parts and determine how parts relate to one another and the overall structure or purpose

- **Evaluate** questions are complex, requiring the test taker to judge based on criteria and standards.

- **Create** problems are the most complex from a cognitive standpoint, requiring test takers to combine elements to form a coherent whole; reorganize into a new pattern or structure.

The knowledge dimension has four categories. **Factual** knowledge is the elementary knowledge a student must be familiar with to understand and solve problems in a subject, **conceptual** if they require students to know of the basic elements in the subject such as principles, generalizations, theories, and models and how these elements relate to one another. Questions are classified as **procedural** if they can be answered algorithmically or technically or **metacognitive** if they test the knowledge of cognition in general as well as awareness and knowledge of one's own cognition.

Depending on the source of the lecture, questions might also have an accompanying answer in the source text, a summary, or their respective learning objectives, that is, an overview of what the student should know after going through a lecture. Despite not being relevant for our current use case, we believe this to be a powerful tool that can be used when training future models on other tasks.

3.3 Dataset Statistics

EduQuest is composed of 68248 questions coming from 76008 lectures. An overview of these is provided in Table 2. Out of these, 162 questions also

	Fact	Pro	MC	Concept
Analyzing	11276	1798	131	1489
Understanding	22170	1503	157	6998
Evaluating	3592	542	409	645
Applying	2995	5424	55	1168
Remembering	1430	344	36	8844
Creating	1169	915	214	156

Table 4: Overview of the cognitive process and knowledge dimensions (Factual - Fact, Procedural - Pro, MetaCognitive - MC and Conceptual - Concept) in *EduQuest*.

have provided answers, and for all questions, the question type is present. Regarding the questions' classifications in the revised Bloom's taxonomy categories, the (Understanding, Factual Knowledge) pair is the most common, followed by (Analyzing, Factual Knowledge), as Table 4 shows. A more comprehensive overview of the relative presence of each question type and Bloom's taxonomy for each data source is provided in the appendix. In general, questions from CK12, OpenText, and OpenStax tend to be simpler and more direct than their counterparts from OCW and Khan Academy, both in their number of words and sentences, as shown in table 6 but also regarding their respective Bloom's taxonomy.

4 Experiments

4.1 Baseline Models

We investigated the suitability of *EduQuest* for training State-of-the-Art (SOTA) Neural Networks on the Question Generation and Summarization tasks. Sample summaries and questions generated by these models are available in the Appendix. Additionally, we provide an online tool through which the reader can input custom text to interact with these models, illustrating their potential in practical applications.

4.1.1 Longformer2Roberta

The Longformer is a natural language processing (NLP) model designed to address the limitations of traditional Transformer models in processing long sequences of text. The Longformer paper introduced an attention mechanism that can scale linearly with sequence length, making it capable of processing much longer sequences. The Longformer has proven to be a significant contribution to the application of Transformer architectures for long document processing and has proven to perform well in various benchmarks. RoBERTa is a variant of the BERT (Bidirectional Encoder Representations from Transformers) mode. RoBERTa differs from BERT in its training methodology, dataset size, which was much larger for RoBERTa, and some hyperparameters. The authors of RoBERTa have shown that it consistently performs well on benchmark tasks. In our experiments, we have used the Longformer as the encoder and RoBERTa as the decoder in an Encoder-Decoder Model for both the question Generation and Summarization task. This model was chosen because it could deal with longer input texts while staying within our hardware limits. The maximum token length was capped at 4096 because of the same hardware limitations. The learning rate (constant at 3e-5) from the Longformer paper was used in training.

4.1.2 T5

T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2019) is an Encoder-Decoder transformer model that reframes all Natural Language Understanding and Natural Language Generation tasks into a unified text-to-text format. T5 uses the original Transformer architecture. Using the large crawled C4 dataset, the model has trained with masked language modeling as well as the SuperGLUE (Wang et al., 2019) tasks by translating all of them to text-to-text tasks.

We experimented with training the T5 base model with 223 million parameters, with the learning rate found to be most promising by the original authors (0.001) as well as a lower learning rate of 0.0001 which improved the performance on the test data from *EduQuest*. The maximum token length was capped at 512 and the learning rate was kept constant during training.

4.1.3 Bloom Lora

LoRA (Low-Rank Adaptation) is a method that accelerates the training of large language models while consuming less memory by freezing pretrained model weights and adding trainable rank decomposition matrices into the model. This technique significantly reduces the number of trainable parameters and storage requirements for task-specific adaptations without introducing inference latency.

Bloom is an autoregressive Large Language Model with 176 billion parameters, Bloom was cre-

	Avg	Std
Summaries	3766.6	1541.2
Questions	368.9	1416.5
Lectures	10174.7	11024.0

(a) Number of Chars

	Avg	StdAvg
Summaries	721.6	307.5
Questions	81.3	315.0
Lectures	2073.9	2640.8

(b) Average Token Length

	Avg	Std
Summaries	5.3	0.3
Questions	4.4	0.8
Lectures	5.0	0.5

(c) Characters per Token

Table 5: General Statistics of the dataset

website	Document Type	Average #Words	Average #Sentences
CK12	Lecture	570.7	37.8
CK12	Questions	12.4	1.1
Khan Acad.	Lecture	45.2	2.4
Khan Acad.	Questions	10.3	1.3
MIT OCW	Lecture	773.5	29.9
MIT OCW	Questions	44.9	3.3
Openstax	Lecture	1625.5	75.3
Openstax	Questions	30.9	2.4
Opentext	Lecture	1096.5	56.7
Opentext	Questions	28.0	2.1

Table 6: Description of lectures and questions extracted from the scraped websites.

ated through a collaborative effort involving over 1,000 researchers and offers a transparent approach to its development and training (Heikkilä, 2022). Bloom's ability to handle a wide range of languages and its open-access nature made it a valuable resource in the field of natural language processing.

We trained a LoRA on Bloom with the following parameters from the paper: We further experimented by increasing the number of attention heads to r=16, which did not improve performance.

4.2 Metrics

We adopt ROUGE (Lin, 2004) and QRelScore (Wang et al., 2022) for the evaluation of QG performance. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a metric commonly used for text summarization and evaluates n-gram overlap, having shown promising results for question generation. QRelScore, on the other hand, is a metric specifically designed to evaluate question generation based on source texts, being able to achieve a higher correlation with human judgments when compared to other metrics. The usage of these two distinct metrics to evaluate the QG models ensures the generated questions are relevant to the source text and also are well-formed and sound natural.

Model	ROUGE	QRelScore
Longformer	0.99	0.99
T5	0.99	0.99
Bloom Lora	0.99	0.99

Table 7: QG benchmarks for Longformer, T5 and Bloom Lora on *EduQuest*

5 Discussion

We present an overview of the results in table 7. Further analysis of the generated questions in Figure (figure with B. Taxonomy) indicates that the models finetuned on *EduQuest* are able to faithfully replicate the cognitive skills acessed by real-life exams and problem sets according to both dimensions of the revised bloom's taxonomy hierarchy.

These results are further supported by qualitatively analysing questions generated by the models before finetuning on *EduQuest* and after training, even for short training times. In fact, the generated questions before training are vague and might not relate well to the lecture content at hand, thus not being suitable to be used in an educational context. However, questions generated by these baselines models after training are significantly more related to the source text, and closely resemble what one

Lecture Text: Government spending and tax rate changes can be useful tools to affect aggregate demand. We will discuss these in greater detail in the Government Budgets and Fiscal Policy chapter and The Impacts of Government Borrowing. **Longformer**: **LongformerQG**: **T5**: **T5QG**: **Bloom Lora**: **Bloom LoraQG**:

Table 8: Qualitative analysis of questions generated by the used baseline models.

would expect from a high-level exam or problem set.

6 Conclusion

In summary, we present *EduQuest*, a large scale dataset for academic question generation. Composed exclusively of expert generated questions, while being the most comprehensive dataset of its kind, consisting of challenging questions across a wide range of subjects. We also show the dataset can be used to generate relevant and high-quality questions from advanced source material.

However, our work has some limitations that warrant future research. Firstly, the majority of the conducted annotations were done using GPT-3.5. Despite yielding promising results, a large-scale human annotation of the question type and bloom's taxonomy evaluation of the presented questions would undeniably prove to be valuable.

We thus believe *EduQuest* will prove to be a valuable tool for the development of new education focused NLP models, and we promote researchers to use it to promote the usage of technology in education.

References

Benjamin S Bloom and David R Krathwohl. 2020. *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. longman.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. In *International Conference on Web and Social Media*.

Xinya Du, Ahmed Hassan, Adam Fourney, Robert Sim, Paul Bennett, and Claire Cardie. 2020. Leveraging structured metadata for improving question answering on the web. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 551–556.

Mellissa Heikkilä. 2022. Inside a radical new project to democratize ai.

Mandar Joshi, Eunsol Choi, Daniel S. 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*, Vancouver, Canada. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Michael J. Prince. 2004. Does active learning work ?

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022. Qrelscore: Better evaluating generated questions with deeper understanding of context-aware relevance.

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