EduQuest: Lecture Texts and Questions for Higher Education

Oliver Holl[0]
ETH Zürich
oholl@ethz.ch

Filipe Szolnoky Cunha[0]
ETH Zürich
fpaesde@ethz.ch

David Streuli
ETH Zürich
dstreuli@ethz.ch

Timothé Laborie
ETH Zürich
tlaborie@ethz.ch

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 Bloom's taxonomy 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.

Keywords

Dataset, Education, Machine Learning, Natural Language Processing

1. INTRODUCTION

Education research has demonstrated that active learning is the most efficient method of learning [16]. In this context, high-quality questions play a crucial role in enabling learners to gauge their understanding of subjects and trigger critical thinking accurately. Moreover, prior research has also established that the number of questions learners receive regarding a specific knowledge concept is directly linked to the efficiency of retaining that knowledge [1], further motivating the need for the generation of complex problems.

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.

There are three main key limitations in existing datasets for educational question generation and summarization that this dataset aims to address: Limited diversity/topics, not being specifically designed for educational purposes and being composed of simple questions that do not require higher-order cognitive skills to solve, only requiring lookups on the text.

The issue of limited diversity is prevalent. Most question-answering datasets, such as SQuAD [18], 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 second limitation concerns the Purpose of Design. Datasets like SQuAD [18], and TriviaQA [10], 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 com-

[0] Equal Contributions.

O. Holl, F. S. Cunha, D. Streuli, and T. Laborie. Eduquest: Lecture texts and questions for higher education. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 849–856, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729970

plexity 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 [3].

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 [18] 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** [10] 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 [11], 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** [15] suffers from similar issues.

LearningQ [5] 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 educational texts created specifically for learning via text, such as textbooks or lecture notes. The vast majority of questions are a collection of the 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 [21] 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

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. A general level comparison between *EduQuest* and other related datasets in shown in Table 1.

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 cover a wide range of subjects, including STEM, social sciences, and others. For *EduQuest*, we filter their books based on their complexity and suitability for high school level and up, available for each OpenStax textbook. Simultaneously we selected texts rich in textual context, and we removed subjects for which the majority of problems were equation-based with little or no textual context, such as Calculus and Algebra. 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

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

[2] https://ed.ted.com/

[3] https://openstax.org/

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

	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

presence of questions at the end of sections, many of the textbooks also include questions at the end of chapters, and spanning the entire textbook, as well as summaries and key terms which 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. 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.

3.1.3 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 Khan Academy lessons and questions were sourced from the LearningQ dataset, which initially did not include instructor-posed questions 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, as we found the learner comments to be often not relevant or of high quality, and have not included or processed them. Orphaned lessons without associated questions were flagged as such but not deleted. Lessons and texts for high school 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, covering the entire MIT curriculum. 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 6529 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, and analyzed the results to make sure they were consistent with the source texts, manually extracting other relevant questions that were missing. GPT-3.5 was used because questions and exercises in OCW were not as easily separable as with the other texts because OCW only provides the questions together in a single document. Given their nature, the scraped questions from OCW are significantly more involved than the ones previously obtained, often having several interlinked subquestions (Appendix B). 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). To ensure the relevance of the CK-12 books for *EduQuest*, a filtering process was applied, retaining only questions suitable for high-school level and above.

3.2 Question Annotation

EduQuest includes meta-features to allow for more variance in the question-generation process. It has been shown [7] 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. **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. **True or False** questions consist of one statement and ask test takers whether that statement is right or wrong regarding the source lecture. **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. **Concept** questions are straightforward, usually requiring the test taker to recall a definition or phrase in the source document. **Open-Ended** questions typically require a longer answer than the other four question types. These questions 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 process-

[4] https://opentextbc.ca/

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

[6] https://ocw.mit.edu/

[7] https://www.ck12.org/student/

ing. 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 has also been classified in the **cognitive process** and **knowledge** dimensions of the revised Bloom's taxonomy [3, 12], 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, with **Understand** questions asking the user to construct meaning from some source text. **Apply** problems, on the other hand, 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. Being more complex, **Evaluate** tasks require the test taker to answer based on criteria and standards. At last, **Create** problems are the most complex from a cognitive standpoint, requiring test takers to combine elements to form a coherent whole and reorganize it into a new pattern or structure.

The knowledge dimension, on the other hand, has four categories. **Factual** (comprising elementary knowledge) **conceptual** (related to principles, theories and models) **procedural** (requiring students to use an algorithmic or technical method) and **metacognitive** (knowledge of cognition). This labeling was done using GPT 3.5 (Appendix B).

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 is a powerful tool for training future models on other tasks. By using these meta-features, fine-tuning models on *EduQuest* allows for customization of the generated questions in a simple manner.

3.3 Dataset Statistics

EduQuest is composed of 68248 questions coming from 76008 lectures. From these, 162 questions also have provided answers, and the question type is present for all questions. Regarding the questions' classifications in the revised Bloom's taxonomy categories, 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 3 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. 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. While flexible training is possible with *EduQuest*, we have trained the models on the combined questions and exercises for each lecture to make training feasible on our hardware limitations. For every lecture text, all available questions were gathered and combined into a single ground truth label up to a maximum of 20 questions. The validation and test sets, respectively, were composed of full books extracted from the OpenStax collection.

Table 2: Overview of the cognitive process and knowledge dimensions (Factual - Fact, Procedural - Pro, MetaCognitive - MC and Conceptual - Concept) in *EduQuest*.

	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 3: Description of the lengths of lectures and questions extracted from the scraped websites by number of words and sentences.

Website	Type	Avg #Words	Avg #Sent.
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

4.1.1 *Longformer2Roberta*

The Longformer [2] is a natural language processing (NLP) model designed to address the limitations of traditional Transformer models in processing long sequences of text, with the original paper introducing an attention mechanism that can scale linearly with sequence length, making it capable of processing much longer sequences. This model 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 [14] is a variant of the BERT (Bidirectional Encoder Representations from Transformers) [6] model. RoBERTa differs from BERT in its training methodology, dataset size, which was much larger for RoBERTa, and some hyperparameters. 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 and the model was trained for ten epochs.

4.1.2 *T5*

T5 (Text-to-Text Transfer Transformer) [17] is an Encoder-Decoder transformer model that reframes all Natural Language Understanding and Natural Language Generation tasks into a unified text-to-text format, that has been trained with masked language modeling as well as the SuperGLUE [19] 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 for ten epochs of training.

4.1.3 Bloom Lora

LoRA (Low-Rank Adaptation) [9] is a method that accelerates the training of large language models while consuming less memory by freezing pre-trained model weights and adding trainable rank decomposition matrices into the model. Bloom [4] is an autoregressive Large Language Model with 176 billion parameters, created through a collaborative effort involving over 1,000 researchers and offering a transparent approach to its development and training [8]. It'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 for one epoch on Bloom with the best performing $r = 4$ as found by the authors. The learning rate was $2e - 4$.

4.2 Metrics

We adopt ROUGE [13] and QRelScore [20] for the evaluation of QG performance. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a widely used metric to evaluate the quality of machine-generated texts by measuring n-gram overlap. QRelScore, on the other hand, is a metric specifically designed to evaluate QG based on source texts, achieving a higher correlation with human judgments than other metrics. Using these two distinct metrics to evaluate the QG models ensures the generated questions are relevant to the source text and well-formed and sound natural. We measured ROUGE-L scores on a test set of 100 lectures with corresponding questions and exercises unseen by the model between the generated questions and the ground truth questions. The QRelScore was computed between the lectures and generated questions of the same test set.

5. RESULTS

We present an overview of the results in Table 4. Further analysis of the generated questions indicates that the models finetuned on *EduQuest* are able to faithfully replicate the cognitive skills accessed by real-life exams and problem sets according to both dimensions of the revised Bloom's taxonomy hierarchy. In particular, this implies that questions that require higher-order cognitive skills, like metacognitive questions, for example, are generated easily using *EduQuest*. These results are further supported by qualitatively analyzing questions generated by the models before finetuning on *EduQuest* and after training, even for short training times, as shown in Table 5. Because the models we've used are text generation or models pre-trained on the summarization tasks, before training, they either didn't generate relevant questions or repetitive text (Bloom/LoRA) or simply summaries (T5 and Longformer2Roberta). However, questions generated by these baseline models after training are significantly more related to the source text and closely resemble what one would expect from a high-level exam or problem set. We note that the Rouge-L Scores improve significantly, while the QRelScores do not, despite the apparent qualitative improvement of the questions. Upon further investigation, this is likely due to QRelScores favoring string overlap, as the official questions also get a significantly lower score. At the same time, a summary consisting of the first two sentences of the lesson texts achieves a consistently high QRelScore. The quantitative results of the summarization tasks can be found in Appendix 7.

Table 4: Rouge and QRrelScores for NLP Models

Model	Rouge-L	QRrelScore
Untrained Models Scores		
T5	15.84	0.089
Longformer2Roberta	15.94	0.253
Trained Models Scores		
T5	18.9	0.084
Longformer2Roberta	28.7	0.048

Table 5: Qualitative analysis of questions generated by the used baseline models. *LongformerQG*, *T5QG*, *Bloom LoraQG* correspond to Longformer, T5 and Bloom Lora after finetuning on *EduQuest*, respectively.

Longformer: Half-life means it will
be with us for a very long time.
Radioactive materials lose some activity
each time a decay event occurs.
LongformerQG: Explain why radioactive decay
occurs Describe the half -life
of a isotope and explain its half-lives
T5: plants sense the length of daylight,
or night length, as a signal to flower.
T5QG:Explain how plants sense seasonal
changes
Explain the relationship of photoperiodism.
Bloom Lora: What is the life cycle of a fern?
What are the different parts of the fern?
What are the different parts of the fern?
Bloom LoraQG: What is the life cycle of
seedless vascular plants?
What is the function of sporangia and
archegonia?

6. CONCLUSION AND FUTURE WORK

In this work, we presented *EduQuest*, a large-scale dataset for academic question generation. *EduQuest* contributes to the field of educational question generation by being the first dataset of its kind composed exclusively of expert-generated questions and lecture texts. We have shown that the dataset can be used to train state-of-the-art language models to generate relevant and high-quality questions from advanced source material. Hence, we believe *EduQuest* to be a valuable contribution to developing new education-focused NLP models, and we are hopeful and excited to see how our colleagues use the dataset and improve it.

7. REFERENCES

[1] H. P. Bahrick, L. E. Bahrick, A. S. Bahrick, and P. E. Bahrick. Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4(5):316–321, 1993.

[2] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[3] B. S. Bloom and D. R. Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain.* longman, 2020.

[4] M. Chandio, S. M. Pandhiani, and R. Iqbal. Article bloom's taxonomy: Improving assessment and teaching-learning process. *Journal of Education and Educational Development*, 3, 01 2017.

[5] G. Chen, J. Yang, C. Hauff, and G.-J. Houben. Learningq: A large-scale dataset for educational question generation. In *International Conference on Web and Social Media*, 2018.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] X. Du, A. Hassan, A. Fourney, R. Sim, P. Bennett, and C. Cardie. 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, 2020.

[8] M. Heikkilä. Inside a radical new project to democratize ai, 2022.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[10] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. 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, July 2017. Association for Computational Linguistics.

[11] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. 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, 2017.

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

[13] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[15] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[16] M. J. Prince. Does active learning work ? 2004.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[19] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019.

[20] X. Wang, B. Liu, S. Tang, and L. Wu. Qrelscore: Better evaluating generated questions with deeper understanding of context-aware relevance, 2022.

[21] Y. Xu, D. Wang, M. Yu, D. Ritchie, B. Yao, T. Wu, Z. Zhang, T. Li, N. Bradford, B. Sun, T. Hoang, Y. Sang, Y. Hou, X. Ma, D. Yang, N. Peng, Z. Yu, and M. Warschauer. 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, May 2022. Association for Computational Linguistics.

APPENDIX

A. QUESTION EXTRACTION

Unlike the other sources, OCW questions were not as easily separable via regex, where every question had a single delimiter, and the formatting was consistent throughout. We thus used the following prompt with the GPT 3.5 to separate the individual questions from OCW exams and assignments:

```
The text delimited by triple backticks
is from a lecture text, your task is
to extract the questions and exercise
prompts without the answers from it.
Format your response in a python list
format.
‘‘‘question‘‘‘
```

To prove the challenge of the MIT OCW source text, consider the following question present in Table 6.

Each of the subquestions in the OCW section of Table 6 are related, but can also be labelled as independent questions on their own. However, we have chosen to label this as one individual question, remaining faithful to the labelling in the original source text. As comparison, we also present example questions from OpenStax and CK12.

B. QUESTION ANNOTATION

We utilized GPT-3.5 to classify questions according to Bloom's revised taxonomy. By engineering a precise prompt for GPT-3.5, we guided the model toward categorizing questions following guidelines. We attempted to label all questions using GPT-3.5. To ensure the reliability of the model's classifications' reliability, we manually reviewed each batch's classifications of 10-40 questions. Effectiveness: Our observations and evaluations indicated that GPT-3.5's classifications were consistent and aligned well with the taxonomy's guidelines and as good as we could have classified them ourselves. However, despite the clear guidelines, these classifications are still subjective, and other human annotators may sometimes disagree.

We acknowledge that we are not educational experts. Therefore, while we are confident in the value of the dataset and the classifications provided by GPT-3.5, we recognize the potential for further validation by educational experts. Although it wasn't feasible for this work to involve educational experts for annotation, a comparative analysis between LLM's classifications and human expert annotations would be a valuable avenue for future research. We provide the used prompt below:

```
‘‘‘Here’s a detailed explanation of the
six levels of the Cognitive Process
Dimension in Bloom’s Revised Taxonomy:

1. Remembering: At this level, students
are required to recall or recognize
information.
This is the most basic level of cognition
and includes simple tasks like memorizing
facts or terms, or retrieving previously
learned material. For example, listing the
capital cities of different countries.
2. Understanding: This level involves
demonstrating an understanding of the
facts, such as interpreting, classifying,
summarizing, inferring, or comparing
information.
Students might explain concepts in their
own words or classify objects into
categories.
For example, explaining the main ideas
of a text.
3. Applying: This level involves using
knowledge in a new situation. It’s about
the practical use of what has been learned,
and may include implementing procedures,
solving problems, or using methods.
For example, using a mathematical formula
to solve a real-world problem.
4. Analyzing: At this level, students
break material into constituent parts,
determine how the parts relate to one
another, and understand the overall
structure.
It includes differentiating, organizing,
and attributing. For example, comparing
and contrasting different economic theories.
5. Evaluating: This level involves making
judgments about the value of material or
methods for given purposes. Students assess
the quality, reliability, or effectiveness
of something, based on certain criteria. For
example, critiquing a piece of literature or
judging the validity of a scientific
experiment.
6. Creating: The highest level of the
taxonomy involves putting elements
together to form a coherent or
functional whole; it’s about
creativity and generating new ideas or
products.
This may include designing, constructing,
planning, or producing. For example, writing
an original research paper or creating a piece
of art.

Additionally, the Revised Bloom’s Taxonomy
incorporates a Knowledge Dimension, which is
orthogonal to the cognitive process dimension,
and it includes four categories:

1. Factual Knowledge: Basic elements essential
for understanding of a discipline.
2. Conceptual Knowledge: Knowledge of
principles, theories, models, classifications,
etc.
3. Procedural Knowledge: Knowledge of how to do
things, methods, techniques, and skills.
4. Metacognitive Knowledge: Knowledge of
cognition in general as well as awareness and
knowledge of one’s own cognition.

Delimited by the triple backticks below are
```

Table 6: Example of an OCW question with several parts, corresponding to the course 14-05, Intermediate Macroeconomics, Spring 2013. Comparison with sample questions from CK12 and OpenStax.

OCW	CK12	OpenStax
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 profits 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]	Why are the underlying economic meanings of the perceived demand curves for a monopolist and monopolistic competitor different? Briefly compare and contrast the incentives found in perfect competition with those found in imperfect competition. Briefly contrast the level that a monopolistically competitive firm will tend to produce at and the price it will charge with that of a perfectly competitive firm.	Define biogeography. Describe how biogeography relates to evolutionary change. Discuss the work of Peter and Rosemary Grant.

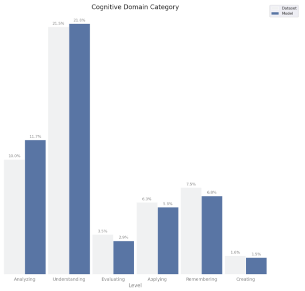

Figure 1: Distribution of Bloom's taxonomy in questions in the original dataset (gray) and from generated questions (LongformerQG) (blue) across the cognitive dimension of Bloom's taxonomy.

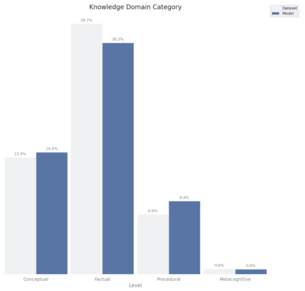

Figure 2: Distribution of Bloom's taxonomy in questions in the original dataset (gray) and from generated questions (LongformerQG) (blue) across the knowledge dimension of Bloom's taxonomy.

```
twenty questions paired with their ids
in this format ‘id: question‘.
Please classify them in Bloom’s revised
taxonomy, first by dimension, then by level,
if a question could belong to multiple
levels, you can add them in a comma separated
string.
Provide your answers in JSON format with the
following keys:
id, dimension, level.‘‘‘{questions}‘‘‘
```

The distributions of the categories for both the generated questions and original questions are shown in figure 2.

C. SUMMARIZATION RESULTS

EduQuest includes 662 Summarizations that are either short summarizations of a specific lesson or summaries of a whole chapter in a textbook of OpenStax. We have trained the Longformer2Roberta (denoted Longformer in the table) and T5 Models with the same parameters as in the QG task on those lecture and summary pairs. It can be noted while T5 does improve, Longformer2Roberta does not see an improvement over the Rouge Metrics suggesting a lot of room for improvements. An overview of the summarization results on *EduQuest* can be found in Table 7.

Table 7: Rouge and QRrelScores for NLP Models

Model	Rouge1	Rouge2	RougeL	RougeLsum
Untrained Models Scores				
T5	15.9	5.5	13.8	14.5
Longformer	41.8	40.3	41.1	41.8
Trained Models Scores				
T5	16.7	5.7	14.0	14.5
Longformer	41.5	40.4	41.0	41.5