EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain

AMIR HADIFAR, SEMERE KIROS BITEW, JOHANNES DELEU,
CHRIS DEVELDER *SENIOR MEMBER, IEEE*, and THOMAS DEMEESTER

Abstract—Natural language processing technology has made significant progress in recent years, fuelled by increasingly powerful general language models. This has also inspired a sizeable body of work targeted specifically towards the educational domain, where the creation of questions (both for assessment and practice) is a laborious/expensive effort. Thus, automatic Question-Generation (QG) solutions have been proposed and studied. Yet, according to a recent survey of the educational QG community's progress, a common baseline dataset unifying multiple domains and question forms (e.g., multiple choice vs. fill-the-gap), including readily available baseline models to compare against, is largely missing. This is the gap we aim to fill with this paper. In particular, we introduce a high-quality dataset in the educational domain, containing over 3,000 entries, comprising (i) multiple-choice questions, (ii) the corresponding answers (including distractors), and (iii) associated passages from the course material used as sources for the questions. Each question is phrased in two forms, normal and cloze (i.e., fill-the-gap), and correct answers are linked to source documents with sentence-level annotations. Thus, our versatile dataset can be used for both question and distractor generation, as well as to explore new challenges such as question format conversion. Furthermore, 903 questions are accompanied by their cognitive complexity level as per Bloom's taxonomy. All questions have been generated by educational experts rather than crowd workers to ensure they are maintaining educational and learning standards. Our analysis and experiments suggest distinguishable differences between our dataset and commonly used ones for question generation for educational purposes. We believe this new dataset can serve as a valuable resource for research and evaluation in the educational domain. The dataset and baselines are made available to support further research in question generation for education (see https://github.com/hadifar/question-generation).

Index Terms—Natural language processing, question generation, multiple-choice questions, transfer learning.

1 Introduction

From the time of Socrates to the present day, questions have been used as an effective teaching technique to facilitate and evaluate comprehension. However, devising high-quality questions has always been challenging and time-consuming due to the extensive human domain knowledge required and the extra effort needed to adapt it to individuals. Davis [1] pointed out that even professional test developers do not manage to write more than three or four good Multiple-Choice Questions (MCQs) per day. Moreover, correction is labor-intensive in a large group setting and may result in delayed feedback to students, especially when multiple graders are involved [2]. Consequently, researchers proposed automatic approaches to facilitate more efficient question construction and correction.

For the *construction* of questions, researchers developed Question Generation (QG) systems. The input of QG systems typically comprises sentences, paragraphs, documents, tables, or images, from which a set of questions is generated. These systems have been used in various ways for educational purposes, such as dialog systems (e.g., Lane and VanLehn [3] suggested a dialog system for novice programmers to help them plan and learn the tacit knowledge of programming) and reading tutor systems (e.g., Alsubait et al. [4] presented a QG system that was designed to help children in grades 1-3 to better understand a text).

For more efficient question *correction*, MCQs were devised. An MCQ contains a stem (i.e., the question itself), the true answer, as well as distractors. The correction process today is relatively straightforward, e.g., using optical mark recognition machines or similar computer vision solutions. As a result, the challenging part that remains to be solved is MCQ generation. The conventional approach to creating an MCQ is to employ a Distractor Generation (DG) model on top of a QG model. To generate distractors, a DG system receives the question, true answer, and possibly source document text [5].

Although automatic MCQ generation systems have been around for a while, most publicly available datasets do not suffice to build systems of sufficient quality for student assessment in the educational domain. Existing MCQ datasets were mainly designed for Question Answering (QA) and created by crowdworkers rather than teachers or educational experts [6]. Teachers' questions typically serve formative and summative assessment needs [7] rather than merely evaluating students' recall skills. Moreover, to achieve trust and adoption by educators, MCQs need to be properly grounded in the source documents — most current MCQ datasets lack such fine- or coarse-grained annotation of source texts.

In this work, we construct a new educational question generation (QG) dataset, EduQG (§3), that contains 3,397 multiple-choice questions (analyzed in detail in §4). The EduQG dataset can be used to finetune existing QG models,

All authors are with the Text-to-Knowledge (T2K) team at IDLab, Ghent University – imec, 9052 Gent, Belgium.
Corresponding author: Amir Hadifar (amir.hadifar@ugent.be)
This work was funded by VLAIO ('Flanders Innovation & Entrepreneurship') in Flanders, Belgium, through the imec-icon *project AIDA ('AI-Driven e-Assessment'), as well as by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.*

A paragraph in a chapter: Physical power—to work the fields, build villages, process raw materials—is a necessity for maintaining a society. During the sixteenth and seventeenth centuries, humans could derive power only from the wind, water, animals, or other humans. Everywhere in the Americas, a crushing demand for labor bedeviled Europeans because there were not enough colonists to perform the work necessary to keep the colonies going. Spain granted encomiendas—legal rights to native labor—to conquistadors who could prove their service to the crown. This system reflected the Spanish view of colonization: the king rewarded successful conquistadors who expanded the empire. Some native peoples who had sided with the conquistadors, like the Tlaxcalan, also gained encomiendas; Malintzin, the Nahua woman who helped Cortés defeat the Mexica, was granted one.

Normal: How could Spaniards obtain encomiendas?

Cloze: Spaniards could obtain encomiendas ___

Options: ✓ **by serving the Spanish crown**
✗ by buying them from other Spaniards
✗ by buying them from native chiefs
✗ by inheriting them

Fig. 1. Example entry of EduQG. For each entry, two formats of the same question have been provided (Normal and Cloze), in addition to the answer key and sentence-level annotation.

as well as benchmarking QG for education (§5 provides baseline models and their performance). As illustrated in Fig. 1, in each MCQ sample, the question is phrased both as a normal question and a cloze form (i.e., fill-the-blank) thereof. This *multi-format* schema is not only valuable from the natural language processing perspective (cf. the link between the cloze form and masked language models) but also for educational purposes [8], [9], [10], [11], [12], [13]. For example, some studies suggested that cloze format questions are preferable to other formats for assessing specific types of knowledge such as reading comprehension and grammar [12], [13]. Also, it was shown that varying formats in assessments induce different approaches to problem-solving and learning [8]. Additionally, we annotated the key sentence(s) in the source text (highlighted in Fig. 1), required to answer the question. Or, inversely, a QG system could in theory generate the question from that key sentence in the context of the source text. Besides, 903 of the questions are accompanied by their cognitive complexity level according to Bloom's taxonomy [14], which facilitates performance analysis of QG and DG models in the function of these question types of varying complexity.

The key contribution of this paper is the composition and analysis of a question dataset in the educational domain, for which the answers are grounded in the source documents. The analysis involves a comparison with available datasets commonly used for question generation (e.g., on the literal overlap of the questions with the answers or the course content). We also provide insights into question generation tasks based on the dataset, in terms of the Bloom's categories of the questions.

The next section summarizes related work in the field of educational question generation and explains how our dataset differs from existing ones. We then describe the dataset collection and annotation process (Section 3), followed by a detailed analysis of our data (Section 4). This is followed by the experimental results on multiple-choice question generation tasks (Section 5). Finally, our conclusions are outlined, with suggestions for future research (Section 6).

2 Related Work

An early notion of QG goes back to the 70s when researchers altered a declarative sentence into an interrogative one by a set of syntactic transformations [15]. They improved this idea by using more sophisticated feature engineering or handcrafted templates. For example, Heilman and Smith [16] used a constituency parser to produce syntactic trees and a set of transformation rules to create questions. Similarly, Kalady et al. [17] included name-entity recognition and keyword extraction to enhance this syntactic-style transformation. Some researchers suggested template-based methods where pre-defined templates are used to generate questions [18], [19]. For instance, Mazidi and Nielsen [19] constructed 50 question templates and utilized a semantic-role-labeling and dependency parser to find the corresponding template for a given input sentence to form a question. While some researchers partially created their own QG dataset [20], thanks to the syntactic nature of the methods, these datasets were mainly used for evaluation purposes rather than training question transformation objectives.

In recent years, there has been a significant shift to sequence-to-sequence models in which a model generates a set of questions given a text segment. As these models depend greatly on training data, having access to a suitable training set is critical. However, due to the unavailability of large-scale training datasets designed for QG and because of the duality of the QG task with QA [21], researchers have adopted the available QA datasets for training QG models [22], [23]. In the next section (§2.1), we will overview some general-purpose QA datasets and then in §2.2 will look into existing datasets in the educational domain.

2.1 Non-Educational question datasets

There is a sizeable body of works (e.g., [24], [25] and the references therein) that focus on QA and QG. The majority of QA datasets are centered around one of these two aspects [24]: (i) *information-seeking* where the questioner did not know the answer, e.g., the questioner submit a query in a search engine to find the answer (ii) *knowledge-probing* where the questioner intended to test the knowledge of another person or machine, e.g., the questioner is a teacher and the answerer is a student.

Two famous examples in the *information-seeking* category are NaturalQ [26] and MS-Marco [27] where questions are generated by internet users and paired with a relevant document(s). Although these datasets have been used for

QG, they are less suitable for the educational setting due to noise and format (users practically never posed a question in multiple-choice or cloze format). The second category, *knowledge-probing*, is more popular among QG researchers with the famous examples of SQuAD [28] and HotpotQA [29]. While the two previously mentioned datasets rely on Wikipedia as a source of information, some relevant researches exist in other domains such as NewsQA [30] which is a QA dataset on news articles, QUASAR-S [31] which is a collection of cloze questions constructed from definitions of software entity tags, or SWAG [32] which is a collection of MCQs, generated from video captions.

These two categories have been utilized for the task of QG, and results suggest that both appear promising for factoid QG. However, their limitations for the educational domain were recently pointed out [6]. These datasets are mainly proposed for testing machines rather than humans, generated by crowdworkers rather than teachers, heavily focusing on name entities and mostly targeting recalling skills. Moreover, their domain was based on Wikipedia or news articles rather than educational textbooks, which is quite different in terms of technical vocabulary, abstract nouns, complex sentences, and ordinary words used in non-typical ways [33].

2.2 Educational question datasets

The above-mentioned limitations tried to address in different ways in the recent past. Researchers proposed datasets that originated from materials written by educators to test students. For example, RACE-C [34], and CLOTH [35] introduced multiple-choice reading comprehension datasets, collected from English examinations. ReClor [36] is multi-choice reading comprehension questions and extracted the data from standardized graduate admission examinations (e.g., GMAT). SciQ [37], TabMCQ [38] and OpenBookQA [39] obtained MCQs from scientific contents, but the questions were written by crowdworkers. It is simply not possible to write a good MCQ in a short limited time [1], [40] which is the case in crowdsourcing annotations (e.g., TabMCQ questions were created in approximately 70 seconds as reported by authors). Moreover, SciQ distractors are generated automatically followed by a post-filtering by crowdworkers which makes it less plausible for students' assessment.

ARC [41] and TQA [42] are collections of MCQs for students. Both datasets only provide question-specific (i.e., answer-key) annotations, and the lack of explicit alignment with course content (e.g., on the sentence or paragraph level), hinders their usage for QG. Similarly, LearningQ [6] introduced a collection of questions where obtained from online teaching platforms. Although this collection contains a large pool of question-document pairs, the absence of fine-grained and coarse-grained answer excerpts only allows for answer-agnostic QG explorations. Closely related to our work, although more narrow in scope (biology questions only) and size (585 questions in total), is ProcessBank [43] which offers a collection of binary choice questions.

The comparison between EduQG and some popular datasets is summarized in Table 1. We divided the table into educational vs. non-educational and compared them regarding: (i) source of creation (*creator*) and question format (*normal* or *cloze*), (ii) its answer type, categorized as Multiple-Choice (*MC*) and/or *extractive* (denoting that spans or entire sentences are considered as answer), and (iii) the *type* and *domain* of their *source text* (e.g., paragraphs vs. entire documents, originating from texbooks or Wikipedia articles).

3 EduQG Development

To develop our EduQG dataset, we chose to work with educational texts and related questions from Openstax,[1] which offers free textbooks and questions that have been developed and peer-reviewed by educators. We crawled all English contents[2] using Openstax's public API, which resulted in a set of 43,578 questions that we then further filtered. Indeed, since we aimed for questions that do not require mathematical reasoning, we excluded[3] topics such as physics, statistics, and algebra. We retained the MCQs, which is a common strategy in available corpora (see the overview in Table 1). This is also motivated by the suitability of MCQs for e-assessment purposes and it has been shown that well-designed MCQs provide a versatile means of assessing a broad range of knowledge, skills, and competencies [1], [44]. Our final EduQG dataset thus amounts to 5,018 questions, which are related to 13 books (283 chapters) and comprise 3,493 *normal* plain text questions and 1,525 in *cloze* form.

We further enriched this crawled question set with (i) grounding passages that support the correct answer (detailed in §3.1), (ii) cloze forms of the plain text questions and vice versa (see §3.2), and (iii) Bloom's taxonomy complexity levels (see §3.3). For (i)–(ii), we employed trained annotators.[4]

3.1 Answer Selection

To support the development of both answer selection and question generation models, we gathered annotations of grounding passage(s) in the source texts for each question. To streamline this process, we automatically retrieve a small list of paragraphs from the relevant chapter, thereby enabling human annotators to efficiently select the correct passages from this list of suggestions rather than undergoing the tedious task of manually searching through the full text. This was implemented in a web platform, with multiple stages: (i) we list the question, and the paragraphs of the corresponding chapter, which the annotator can choose to view either in the original order or ranked by BM25 [45] (using as a query the concatenation of question and correct answer); (ii) then the annotator selects a paragraph that fully/partially contains information leading to the correct answer; (iii) for a selected paragraph, the annotator subsequently selects one or more (and no more than strictly needed) sentences comprising the answer information; (iv) we repeat steps (ii)–(iii) to allow the

1. https://openstax.org
2. Available under a Creative Commons License.
3. Appendix A lists all Openstax books that we retained.
4. We hired two master's students in linguistics with prior experience in annotation projects, compensated at 13.5 EUR/h.

TABLE 1
Qualitative comparison of datasets.

	Dataset	question			answer		source text	
		creator	normal	cloze	MC	extractive	type	domain
Non-educational	SQuAD	crowd	✓	-	-	✓	paragraph	Wikipedia
	HotpotQA	crowd	✓	-	-	✓	paragraph	Wikipedia
	NaturalQ	crowd	✓	-	-	✓	document	Wikipedia
	NewsQA	crowd	✓	-	-	✓	document	News
	SWAG	crowd	✓	-	✓	-	-	Video captions
	QUASAR-S	auto	-	✓	-	✓	paragraph	Stackoverflow
Educational	CLOTH	educator	-	✓	✓	✓	paragraph	Standardized tests
	RACE-C	educator	-[1]	✓	✓	✓	paragraph	Standardized tests
	ReClor	educator	✓	-	✓	-	paragraph	Standardized tests
	LearningQ	educator	✓	-	-	-	document	Online courses
	ARC	educator	✓	-	✓	-	-	textbook
	OpenBookQA	crowd	✓	-	✓	✓	paragraph	textbook
	TQA	educator	✓	-	✓	-	document	textbook
	SciQ	crowd	✓	-	✓	✓[2]	paragraph	textbook
	ProcessBank	educator	✓	-	✓	✓	paragraph	textbook
	EduQG *(Ours)*	educator	✓	✓	✓	✓	document	textbook

[1] If questions are either *normal* or *cloze* form, the majority is chosen as representative of a column.
[2] SciQ provides answers in the form of paragraphs rather than spans or sentences.

annotator to indicate multiple supporting passages (for at most 15 minutes per question). To make the pipeline more clear, further details, including screenshots of the annotation interface, are provided in Appendix B.

We further applied a new post-processing step to check the questions that were flagged by the annotators as problematic, and filter them out if necessary. The latter questions mainly fall into four categories: 1) arithmetic questions (e.g., requiring mathematical reasoning) 2) non-factoid questions (e.g., requiring a complex answer such as opinion) 3) media-related questions (e.g., requiring a plot to answer the question). 4) un-answerable questions (e.g., annotators could not find the answer in the given time). The retained collection after post-processing contains 3,397 MCQs, of which 1,356 are in *cloze* form and 2,041 in *normal* form.

To measure annotation agreement we randomly sampled 125 questions and assigned them to two annotators. For the higher level annotation (i.e., support/partial vs. no-support), annotators display an agreement of 90.4% of the paragraphs, with a Kappa score of 0.8. For the finer level annotation (i.e., sentence selection), annotators fully agreed on selecting the same set of sentences for 42.9% of the paragraphs, while agreement on at least one common sentence between two selections (i.e., full or partial agreement) is 76.3%. For 23.7% of paragraphs, we found no agreement between annotators. We hypothesize that this is due to the fact that frequently multiple distinct (sets of) sentences in a chapter allow answering the same question. It should be noted that calculating Kappa scores for comparing the sentence selection appeared less straightforward, and therefore we decided to only report the agreement percentage instead.

3.2 Question Generation

In the second phase of the annotation process, we further enriched the dataset. For each question, we created a semantically identical but structurally different counterpart. More specifically, we hired two linguistic experts to write a *cloze* formulation of a given MCQ from its *normal* formulation. This conversion not only renders the dataset more homogeneous; it also gives us the opportunity to tackle the new task of question format conversion (see §5.3), with potential practical value for e-assessment systems. The experts converted a *normal* MCQ to its corresponding *cloze* formulation by considering two rules: i) The gap is replaceable by all candidate answers, which then leads to a grammatically correct and meaningful sentence. ii) No information is added or left out, compared to the original question. Therefore, we asked them to use the same phrases, tenses, etc., as much as possible. The same annotation process was repeated in the opposite direction. We showed each *cloze* MCQ to the experts, asking them to convert it to the *normal* formulation. This step added 1,356 new *normal* MCQs and 2,041 new *cloze* formulations, which can be considered equally educationally valuable as their original teacher-generated counterparts.

3.3 Bloom's Taxonomy Labels

The Openstax's API offers access to the revised Bloom taxonomy [14] for some questions. The Bloom taxonomy is one of the most recognized cognitive schemes for classifying questions into different levels of complexity, and it is widely used in the development of test items in the educational community [46], [47]. It categorizes questions into six increasing levels of complexity: *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*. Lower levels (e.g., *Remember*) are suitable for assessing students' preparation and

comprehension or for reviewing and summarizing content, while higher levels (e.g., *Create*) encourage students to think critically and to solve problems [1]. Among all retained questions after filtering the dataset (see §3.1) we were able to further enrich 903 questions with the Bloom's taxonomy label. The distribution of questions among the different levels is as follows: 660 in *Remember*, 114 in *Understand*, 110 in *Apply*, and 19 in *Analysis*. No questions for the last two categories (*Evaluate* and *Create*) were found, likely due to the fact that we dropped mathematical questions, hypothetical, and opinion-based questions. Although this might seem as a limitation in our study, it is important to note these two categories are less prevalent in current-day factoid-based QG.

3.4 EduQG Statistics

Table 2 presents some statistics of EduQG. The first four rows show the statistics of entry pairs (*normal* and *cloze*), distractors, chapters, and courses in EduQG. The next row (5) counts the questions of the types None-Of-The-Above (NOTA) or All-Of-The-Above (AOTA). We brought this feature up because, in addition to the answer's length, it plays an important role in the focusability of MCQs (an MCQ should be answerable without looking at the response options). Although the focusability has received almost no attention from the QG community, it has its own place in educational and teaching research [48]. Rows 6 to 8 present the average number of words in answers, questions, and chapters. Rows 9-13 list the number of questions per Bloom label.

TABLE 2
Statistics of EduQG.

Feature	Statistic
1. # (*normal*, *cloze*) pairs	3397
2. # of distractors	10172
3. # of chapters	241
4. # of courses	12
5. # of NOTA and AOTA	92
6. Avg. length of answers	4.1
7. Avg. length of questions	12.3
8. Avg. length of chapters	12641.5
9. # of pairs w Bloom's taxonomy	903
10. *Remember*	660
11. *Understand*	114
12. *Apply*	110
13. *Analysis*	19

4 Data Analysis

We analyze EduQG in terms of four criteria: the minimum required sentences, question or answer overlap with the minimum required sentences, as well as Bloom's labels, and compare it with SQuAD (Table 3). We chose to use SQuAD not only because it is the most commonly used dataset for QG but also one of the very few datasets where questions and grounding answers are aligned. None of the educational datasets (to our knowledge) provide such a feature. This alignment limits the variability in generated questions and therefore leads to more accurate automatic evaluation metrics when compared with the teachers' questions.

The **Minimum Required Sentences**, denoted by MRS in the table, is one of the factors that affect the difficulty of QG and we define the minimum required sentences as the smallest set of reference sentence(s) in the available course material that allows a human to answer the considered question.[5] We assume it corresponds to the sentences selected by the annotators within the grounding paragraphs (see §3.1). The number of required sentences (i.e., MRS) varies from a *single* sentence to *multiple* ones. The questions that rely on multiple sentences appear more thought-provoking to create, compared to the single-sentence ones. In our collection, answering 37.6 percent of the questions relies on a *single* sentence, and 62.4 percent on *multiple* one. In SQuAD, all answers are short spans literally mentioned in a *single* sentence, which means its MRS is 100 percent for *single* cases (see Table 3).

The **Answer-MRS overlap** is quantified as the percentage of questions for which the answer literally appears in the MRS. Higher values indicate students may more easily select the correct answer through memorization. We calculated the lexical match in two ways: *exact* vs. *normalized* (the latter referring to lowercasing, stopword and punctuation removal, stemming, on both answer and MRS). As shown in the table (under Answer-MRS overlap), 38.1 percent of the answers exactly appeared on the MRS in our dataset, and this number increases to 53.9 when we apply the normalization. For SQuAD, this number is 100 percent since by construction, all answers are literally mentioned in the MRS.

The **Question-MRS overlap** is measured as well, assuming that real-world educational questions typically transcend a simple syntactic transformation of a declarative sentence [49]. The third block in the table (Question-MRS overlap) presents the results for different levels of *normalized* word overlap (with the same normalization steps as for the normalized Answer-MRS overlap) between the question and MRS (i.e., the number of words they have in common after normalization). For example, 14.5 percent of questions in our dataset have at most one word in common ($0 \leq overlap \leq 1$), compared to is 12.1 percent in SQuAD. Although the gap might seem negligible, we should consider that 62.4 percent of the questions in EduQG rely on multiple sentences. The results for high-overlapping cases (e.g., $5 < overlap$) also indicate a slightly higher tendency of crowdworkers to reuse the same words from the context compared to the educators.

Bloom's labels can be a valid proxy to estimate the cognitive level required to answer the question. Therefore this factor can directly contribute to the QG difficulty [50]. Not only from the QG perspective but also for learning purposes that is important, since teachers usually combine a mixture of easy and difficult questions to differentiate between weaker or stronger students in the subject. Therefore a suitable educational dataset should be representative of the variant cognitive levels. As presented in the table, most questions (73.1%, according to the subset of Bloom-labeled

5. We do not take into account background knowledge, common reasoning, etc., that the annotators relied on to judge answerability in combination with the selected questions.

questions) in our dataset fall into the *Remember* category, and the rest goes into the next three levels. Although that seems out of balance at a first glance, it is in line with literature indicating that about 70 percent of asked questions are shallow, and the rest are deep and high-level ones [51]. For SQuAD, all questions reside in the first level as stated in [6].

TABLE 3
The comparison between our collection and SQuAD with regarding four criteria: number of minimum required sentences (MRS), Answer-MRS overlap, Question-MRS overlap, and Bloom's labels. The MRS stands for the minimum number of sentence(s) needed to answer a question.

Property	Type	EduQG	SQuAD
MRS	*Single*	37.6	100
	Multiple	62.4	0.0
Answer-MRS overlap	*Exact*	38.1	100
	Normalized	53.9	100
Question-MRS overlap	$0 \leq overlap \leq 1$	14.5	12.1
	$2 \leq overlap \leq 3$	45.8	41.1
	$4 \leq overlap \leq 5$	27.0	31.8
	$5 < overlap$	12.7	15.0
Bloom's label	*Remember*	73.1	100
	Understand	12.6	0.0
	Apply	12.2	0.0
	Analyze	2.1	0.0

5 MCQ Experiments

We now present a number of experiments, with a focus on the creation of MCQ (although the EduQG dataset could be used for the task of QA as well). In particular, we cover the tasks of question generation (§5.1), distractor generation (§5.2), and question format conversion (§5.3). The first two experiments are necessary components for developing multiple-choice questions. The last experiment is on question format conversion, which is optional for MCQ, although it may help instructors to switch formats at their wish. The provided baselines serve as a benchmark for further research towards tools for educators to quickly and easily generate high-quality MCQs for use in their classes, assessments, and other educational materials (see Appendix C for ethical considerations). Furthermore, our question complexity section (§5.4) provides insight into the relationship between question difficulty and model effectiveness as quantified by popular evaluation metrics.

For all experiments, we sampled 20% of chapters along with their questions for testing. We preferred the chapter-level split over randomly sampling questions to avoid unwanted and indirect knowledge transfer across questions in the chapter. This division leads to 671 questions for testing and 2,726 for training. Among 671 questions in the test-set, 176 questions have Bloom's taxonomy labels that are distributed as follows: 136 in *Remember*, 18 in *Understand*, 18 in *Apply*, and 4 in *Analyze*. For all generation models, we adopt simple but strong baselines by finetuning the 'base' version of T5 [52] on the target tasks. Further details about the experimental settings and hyperparameters are reported in Appendix D. For the evaluation, we employed standard metrics to evaluate the results: BLEU [53], ROUGE$_l$ [54], METEOR [55], token-level F1-score [28], and Exact Match (EM).

5.1 Question Generation

We first investigate the potential of finetuning T5 [52] for educational QG. Three fine-tuned versions are evaluated on the EduQG test-set: (i) T5 finetuned on SQuAD, (ii) T5 finetuned on EduQG, (iii) and T5 multi-stage finetuned, e.g., first finetuned on SQuAD, and then on EduQG (SQuAD→EduQG). Different from the standard finetuning paradigm, in multi-stage finetuning, the pretrained model is further trained on multiple related tasks in stages. The idea is to gradually adapt the pretrained model to the specific requirements of the new task(s) by training on smaller yet more focused subsets of data until the final, fine-tuned model is obtained. This approach allows to transfer of the knowledge learned on related tasks [56], [57], [58]. Thus, T5 was initially finetuned on SQuAD (large, with general-purpose questions) for 10 epochs, and then finetuned on EduQG (smaller, but with education-oriented questions) for another 10 epochs. For all variations, we finetuned T5 in an answer-aware mode where the model receives not only the context but also the target answer as input to generate the question. Although leaving out the answer may seem more appealing, we observed that it is more likely to result in unanswerable questions [59]. The context and answer are concatenated, separated by a special token `[SEP]`, and the output is decoded using a greedy search schema.

The first block in Table 4 presents our QG results. The numbers are in line with earlier studies that reported multi-stage finetuning is beneficial and in-domain finetuning (denoted by EduQG) can boost the performance [56], [57]. In particular, the multi-stage finetuning (SQuAD→EduQG) consistently outperforms the two other variants on all evaluation metrics (e.g., by 6.86 and 0.79 BLEU points over SQuAD and EduQG, respectively). Importantly, in-domain finetuning on EduQG is superior to finetuning on the much larger SQuAD dataset. This confirms the complementary value for QG of the new EduQG dataset for the educational domain.

We followed this conventional paradigm to evaluate the quality of generated questions since other techniques, such as asking teachers to review the quality of generated questions or connecting question quality to student performance, are not reproducible. Alternatively, asking crowdworkers to rate the quality of questions would lead to potential new biases [60]. Another possible assessment technique is to replace human responses with an artificial crowd of question-answering models. The validity of such an approach would however be extremely reliant on the type of questions and the specific domain.

5.2 Distractor Generation

We next look into the automatic generation of distractors. We follow the same paradigm as in §5.1 to build our baselines. Different variants of T5 devised to generate distractors include: (i) T5 finetuned on RACE [61], (ii) T5 finetuned on EduQG, and (iii) T5 multi-stage finetuned, first on RACE and then EduQG (RACE→EduQG). We used parallel decoding, i.e., with all required distractors per question generated

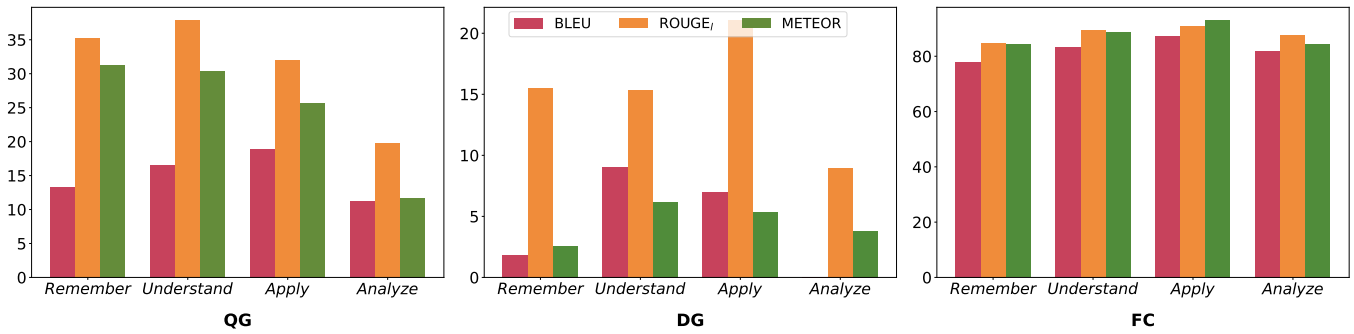

Fig. 2. The performance of our models on Question Generation (QG), Distractor Generation (DG), and Format Conversion (FC) for different levels of Bloom's taxonomy.

TABLE 4
MCQ experiments evaluated on the EduQG test set. For question generation (QG) and distractor generation (DG), we finetuned T5 in various ways including out-of-domain finetuning (e.g., on SQuAD or RACE), in-domain finetuning (e.g., EduQG), and multi-stage finetuning (e.g., SQuAD→EduQG or RACE→EduQG). For the format conversion (FC) experiment, two strategies are employed: single-prompt, with T5 finetuned in one direction (e.g., cloze→normal), and multi-prompt with T5 finetuned in both directions but tested on the targeted (underlined) format (e.g., cloze↔normal indicates evaluation on the normal format of the test instances).

Experiment	Model	BLEU	ROUGE$_l$	METEOR	EM	F1
QG	SQuAD	8.55	24.78	20.08	0.0	25.81
	EduQG	14.62	33.38	28.70	0.45	35.12
	SQuAD→EduQG	**15.41**	**34.26**	**29.65**	**0.75**	**36.29**
DG	RACE	10.88	27.50	18.10	7.28	21.45
	EduQG	17.71	32.53	20.39	9.32	24.33
	RACE→EduQG	**17.73**	**34.13**	**21.54**	**10.48**	**25.74**
FC	cloze→normal	80.01	86.25	86.71	40.39	89.46
	cloze↔normal	80.26	86.38	86.95	40.69	89.71
	normal→cloze	79.97	90.05	87.20	52.46	93.05
	normal↔cloze	80.24	90.36	87.36	53.80	93.17

in a single decoding step as a `[SEP]`-separated list. Note that we did not evaluate the diversity (or lack thereof) of the generated distractors for the reported baselines. The results reported in the second block of Table 4 are averaged metrics for the different generated distractors, each time evaluated against the entire set of ground truth distractors.

As for the QG task, multi-stage finetuning allows improving upon in-domain finetuning. In-domain finetuning, in turn, is superior to the RACE-only model.

5.3 Question Format Conversion

The multi-format property of EduQG invites new prediction tasks, and in particular Format Conversion (FC), i.e., the automatic conversion between question formats, in particular from *normal* to *cloze* or vice versa. On the one hand, this task is a desirable feature for developers of educational tests. On the other hand, it could be useful for evaluating masked language models [62], and it could support the creation of more challenging datasets [63].

Again, we propose different strategies to finetune T5 for this task, with results reported in the third block of Table 4. First, we finetuned T5 in a single-prompt fashion, based on one input format, and converted it into the other (e.g., 'cloze→normal'). Next, in a multiple-prompt experiment, we finetuned conversion in both directions simultaneously. Test results are reported separately (with the underlined format indicating the evaluation format, as in 'cloze↔normal' for evaluation on the normal format of the test instances), to allow for comparison with the corresponding single-prompt experiment. The multiple-prompt model slightly but consistently outperforms single-prompt finetuning on all evaluation metrics (see generated samples in Appendix E).

5.4 Question Complexity

We further investigated the baseline models in terms of question complexity. We selected the best-performing models in previous experiments and evaluated their performance on the Bloom-annotated subset of the test questions. As can be seen from Fig. 2, increasing levels of difficulty do not necessarily lead to decreased performance for all metrics. For example, the BLEU score for QG as well as DG seems to increase rather than decrease in going from level 1 (*remember*) to level 2 (*understand*). There is a clear decrease in level 4 (*analyze*), though, indicating that the model struggles to compete with human-generated questions and distractors of that level. This may suggest that the current automatically generated questions may not be able to assess students' deeper understanding of the material and only test their ability on surface-level information. We also notice that

the format conversion results are rather indifferent to the difficulty levels, with high scores for all evaluation metrics.

6 Conclusions

In this paper, we introduced EduQG, a new dataset for educational QG based on a collection of OpenStax textbooks, whereby course contents and questions are generated by educational experts. The dataset offers 3,397 high-quality multiple-choice questions, each phrased in a *cloze* as well as *normal* form, and its corresponding answer is linked to the relevant chapter text. Moreover, 903 questions are linked to their cognitive complexity according to Bloom's taxonomy. We analyzed the data and provided baseline results for question generation, distractor generation, and question format conversion, with clear added value w.r.t. out-of-domain datasets such as RACE and SQuAD. The baseline results indicate that there is still much room for improvement when it comes to educational QG, and we believe that the future of this field lies in expanding the dataset to cover a wider range of question formats (e.g., True/False or Open-ended), question types (e.g., multi-modal), and question complexities (e.g., *evaluate* or *create*). We hope EduQG will stimulate the development of more advanced teacher-assistant models.

Acknowledgment

We would like to thank the project partners, Televic Education and WeZooz Academy, for contributing data and use cases.

Appendix A
List of Study Books

The following is a list of the books we used as the data source for EduQG:

1) American Government. https://openstax.org/details/books/american-government-2e
2) Anatomy and Physiology. https://openstax.org/details/books/anatomy-and-physiology
3) Biology. https://openstax.org/details/books/biology-2e
4) Business ethics. https://openstax.org/details/books/business-ethics
5) Business Law i Essentials. https://openstax.org/details/books/business-law-i-essentials
6) Intellectual Property. https://openstax.org/details/books/introduction-intellectual-property
7) Introduction to Sociology. https://openstax.org/details/books/introduction-sociology-2e
8) Microbiology. https://openstax.org/details/books/microbiology
9) Financial Accounting. https://openstax.org/details/books/principles-financial-accounting
10) Managerial Accounting. https://openstax.org/details/books/principles-managerial-accounting
11) Psychology. https://openstax.org/details/books/psychology-2e
12) U.S. History. https://openstax.org/details/books/us-history

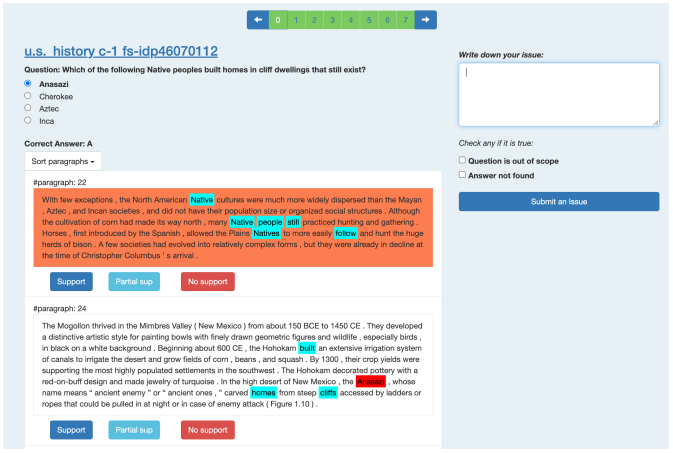

Fig. 3. A screenshot of the first stage in our annotation platform. A question, a list of options, and aligned paragraphs are shown to the annotator. The annotator should select the label among support, partial-sup, or no-support.

Appendix B
Annotation Platform

In this section, we provide some screenshots of the annotation platform. Fig. 3 shows the first stage in the platform. The question, the options (answer in boldface), and a list of paragraphs are shown to the annotator. For each paragraph, three buttons are designed (Support, Partial-support, and No-support). In case of any issues, the annotator can submit it in the info box (on the right side of the figure). Fig. 4 depicts the second stage of annotation. The annotator enters the answer selection process after choosing (Support or Partial-support). The selected paragraph, along with the question and options, is shown to the annotator, and they should select the relevant sentence(s) that answer the question.

Appendix C
Ethical Considerations

We assume the dataset and models that were introduced through this research are low-risk in terms of potential harm to individual people. The dataset is a selection from existing (manually created) educational content, enriched with meta-data (between questions and course material), and we are convinced our compilation of the dataset has not induced any additional ethical risks. However, when training question generation models as the ones we benchmarked in this paper, and using these for educational purposes, we want to stress that there is a need to ensure accountability, and to establish clear guidelines for their deployment. Being derived from general-purpose neural language encoders that have been trained on real-world and therefore potentially biased or discriminatory content [64], our models may have inherited some of these properties, and could therefore generate similarly biased text. Therefore, it is important for educators and researchers to carefully consider these ethical issues and ensure that the generated questions are aligned with educational goals and do not perpetuate harmful biases. Educators should have

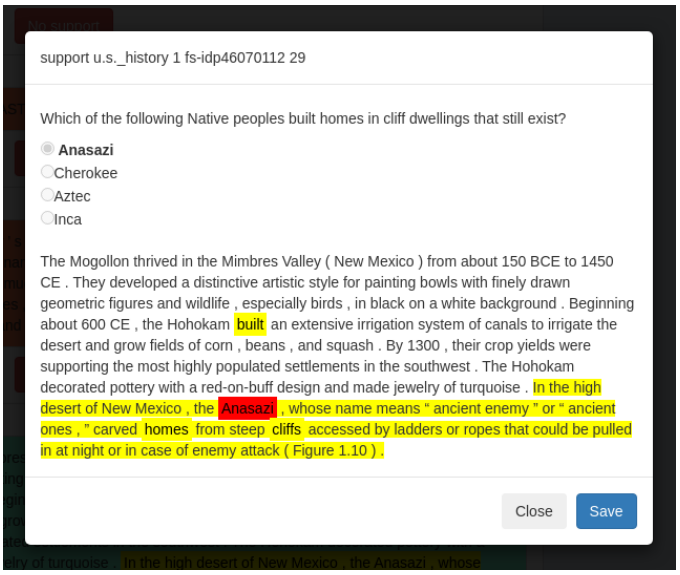

Fig. 4. A screenshot of the second stage in our annotation platform. After selecting the paragraph label, the question, the options, and the paragraph itself are shown to the annotator for sentence selection. The annotator should select relevant sentence(s) that supports answering the question.

the final say in accepting or modifying question suggestions generated by such models, with their educational goals in mind (e.g., in terms of formative and especially summative assessment). In practice, these models are meant to help increase teachers' efficiency in preparing teaching materials, rather than replacing teachers in any way (which is out of the question, given the difficulties in educational question generation, as demonstrated in this work). A key benefit of a higher efficiency in AI-supported manual question generation is the possibility for personalized approaches towards students.

Note that the EduQG collection was gathered from publicly available resources provided by Openstax.[6] The books are freely available online under Creative Commons (CC by 4.0) licenses which means one can distribute the material under a different license elsewhere, publish it for profit, or transform it without endangering the original Openstax repository.

Appendix D
Empirical Setup

All experiments were implemented in Pytorch [65] and were run with one GPU (NVIDIA Tesla V100) with CUDA 10.1 with full precision (FP32) training. T5 was finetuned from the pretrained 'base' version of the huggingface[7] library. We used the following hyperparameters for the QG and FC experiments:

```
batch size =  8
total epochs =  10
optimizer = AdamW
weight decay = 0.1
adam epsilon = 1e-08
max grad norm = 1.0
lr scheduler = linear
warmup steps = 500
max source length = 512
max target length = 48
learning rate = 5e-5
gradient accumulation steps = 4
```

For the distractor generation experiment, a smaller learning rate (1e-5) and more training epochs (20) were used. The maximum target length was also increased to 150 in order to avoid truncation of the outputs.

Appendix E
Example Predictions

Table 5 shows cherry-picked examples of our best-performing models for each of the experiments in §5. Each block in the table corresponds to a specific experiment and the corresponding best-performing model. Each example displays the model's input (Input), the ground truth from a textbook (Ground), and generated output by the model (Prediction). These examples showcase the models' ability to generate coherent and semantically meaningful outputs.

6. https://github.com/openstax
7. https://huggingface.co/

References

[1] B. G. Davis, *Tools for teaching*. John Wiley & Sons, 2009.
[2] M. K. Kim, R. A. Patel, J. A. Uchizono, and L. Beck, "Incorporation of bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course," *Amer. J. Pharmaceutical Educ.*, vol. 76, no. 6, 2012.
[3] H. C. Lane and K. VanLehn, "Teaching the tacit knowledge of programming to novices with natural language tutoring," *Comput. Sci. Educ.*, vol. 15, no. 3, pp. 183–201, 2005.
[4] T. Alsubait, B. Parsia, and U. Sattler, "Ontology-based multiple choice question generation," *Künstliche Intelligenz*, vol. 30, pp. 183–188, 2016.
[5] Y. Gao, L. Bing, P. Li, I. King, and M. R. Lyu, "Generating distractors for reading comprehension questions from real examinations," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, 2019, pp. 6423–6430.
[6] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, "Learningq: A large-scale dataset for educational question generation," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, 2018.
[7] G. Marbach-Ad and P. G. Sokolove, "Can undergraduate biology students learn to ask higher level questions?" *J. Res. Sci. Teaching*, vol. 37, no. 8, pp. 854–870, 2000.
[8] J. H. Holley and E. K. Jenkins, "The relationship between student learning style and performance on various test question formats," *J. Educational Bus.*, vol. 68, no. 5, pp. 301–308, 1993.
[9] S. F. Reardon, D. Kalogrides, E. M. Fahle, A. Podolsky, and R. C. Zárate, "The relationship between test item format and gender achievement gaps on math and ela tests in fourth and eighth grades," *Educational Researcher*, vol. 47, no. 5, pp. 284–294, 2018.
[10] K. F. Stanger-Hall, "Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes," *CBE Life Sci. Educ.*, vol. 11, no. 3, pp. 294–306, 2012.
[11] M. E. Martinez, "Cognition and the question of test item format," *Educational Psychologist*, vol. 34, no. 4, pp. 207–218, 1999.
[12] R. B. Baldauf Jr and I. K. Propst Jr, "Matching and multiple-choice cloze tests," *J. Educational Res.*, vol. 72, no. 6, pp. 321–326, 1979.
[13] A. S. Gellert and C. Elbro, "Cloze tests may be quick, but are they dirty? development and preliminary validation of a cloze test of reading comprehension," *J. Psychoeducational Assessment*, vol. 31, no. 1, pp. 16–28, 2013.
[14] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

TABLE 5
Some generated examples from the best performing model in question generation (QG), distractor generation (DG), and format conversion (FC) experiments.

Experiment	Example
QG	**Input**: competition [SEP] According to sociologist and philosopher Karl Marx, such conflicts are necessary, albeit ugly, steps toward a more egalitarian society. In this pattern of change through conflict, people tend to gain greater personal freedom and economic stability (1848). Philosopher and social scientist Karl Marx was a seminal force in developing the conflict theory perspective. He was a proponent of conflict , in general , because he felt that it was the only means of promoting positive change for the underprivileged. Marx did not agree with Durkheim's notions of cooperation and interdependence; he instead saw society as a stage for exploitation and strife. **Ground**: Karl Marx believed social structures evolve out of: **Prediction**: Marx believed that society should be a stage for what?
DG	**Input**: Adams , Murray , and Warren all came from privileged backgrounds . All three were fully literate , while many women in the American republic were not . Their literacy and station allowed them to push for new roles for women in the atmosphere of unique possibility created by the Revolution and its promise of change . Female authors who published their work provide evidence of how women in the era of the American Revolution challenged traditional gender roles . Inspired by the Revolution , Judith Sargent Murray of Massachusetts advocated women 2019 s economic independence and equal educational opportunities for men and women (Figure 7.5) . Murray , who came from a well-to-do family in Gloucester , questioned why boys were given access to education as a birthright while girls had very limited educational opportunities . She began to publish her ideas about educational equality beginning in the 1780s , arguing that God had made the minds of women and men equal . Another privileged member of the revolutionary generation , Mercy Otis Warren , also challenged gender assumptions and traditions during the revolutionary era (Figure 7.5) . Born in Massachusetts , Warren actively opposed British reform measures before the outbreak of fighting in 1775 by publishing anti-British works . In 1812 , she published a three-volume history of the Revolution , a project she had started in the late 1770s . By publishing her work , Warren stepped out of the female sphere and into the otherwise male-dominated sphere of public life . Some women hoped to overturn coverture . From her home in Braintree , Massachusetts , Abigail Adams (Figure 7.4) wrote to her husband , Whig leader John Adams , in 1776 [SEP] Which of the following figures did not actively challenge the status of women in the early American republic? [SEP] phillis wheatley **Ground**: [”abigail adams”, ”mercy otis warren”, ”judith sargent murray”] **Prediction**:[”john adams”, ”judith sargent murray”, ”mercy otis warren”]
FC	**Input**: What the electrons added to NAD+ do, is that ___. **Ground**: What do the electrons added to NAD+ do? **Prediction**: What do the electrons added to NAD+ do? **Input**: How many NADH molecules are produced on each turn of the citric acid cycle? **Ground**: The number of NADH molecules that are produced on each turn of the citric acid cycle is ___. **Prediction**: On each turn of the citric acid cycle, ___ NADH molecules are produced.

[15] J. H. Wolfe, "Automatic question generation from text - an aid to independent study," in *Proc. Spec. Interest Group Comput. Sci. Educ. (SIGCSE)*, 1976, pp. 104–112.

[16] M. Heilman and N. A. Smith, "Question generation via overgenerating transformations and ranking," Carnegie Mellon Univ. Lang. Technol. Inst., Tech. Rep., 2009.

[17] S. Kalady, A. Elikkottil, and R. Das, "Natural language question generation using syntax and keywords," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2010, pp. 376–385.

[18] J. Mostow and W. Chen, "Generating instruction automatically for the reading strategy of self-questioning." in *Proc. Int. Conf. Artif. Intell. Educ. (AIED)*, 2009.

[19] K. Mazidi and R. D. Nielsen, "Leveraging multiple views of text for automatic question generation," in *Proc. Int. Conf. Artif. Intell. Educ. (AIED)*, 2015, pp. 257–266.

[20] N. A. Smith, M. Heilman, and R. Hwa, "Question generation as a competitive undergraduate course project," in *Proc. NSF Workshop Question Gener. Shared Task Eval. Challenge*, 2008.

[21] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2017, pp. 866–874.

[22] D. Su, Y. Xu, W. Dai, Z. Ji, T. Yu, and P. Fung, "Multi-hop question generation with graph convolutional network," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4636–4647.

[23] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, and R. G. Baraniuk, "Qg-net: A data-driven question generation model for educational content," in *Proc. Annu. ACM Conf. Learn. Scale (L@S)*, 2018, pp. 1–10.

[24] A. Rogers, M. Gardner, and I. Augenstein, "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension," *ACM Comput. Surv.*, vol. 55, no. 10, pp. 1–45, 2022.

[25] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, pp. 121–204, 2020.

[26] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 452–466, 2019.

[27] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human generated machine reading comprehension dataset," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2016, p. 660.

[28] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2016, pp. 2383–2392.

[29] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable

multi-hop question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 2369–2380.

[30] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," in *Proc. Workshop Representation Learn. Natural Lang. Process. (Rep4NLP)*, 2016, pp. 191–200.

[31] B. Dhingra, K. Mazaitis, and W. W. Cohen, "Quasar: Datasets for question answering by search and reading," *arXiv:1707.03904*, 2017.

[32] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 93–104.

[33] C. A. Espin, T. W. Busch, E. S. Lembke, D. D. Hampton, K. Seo, and B. A. Zukowski, "Curriculum-based measurement in science learning: Vocabulary-matching as an indicator of performance and progress," *Assessment Effective Intervention*, vol. 38, no. 4, pp. 203–213, 2013.

[34] Y. Liang, J. Li, and J. Yin, "A new multi-choice reading comprehension dataset for curriculum learning," in *Proc. Asian Conf. Mach. Learn. (ACML)*, 2019.

[35] Q. Xie, G. Lai, Z. Dai, and E. Hovy, "Large-scale cloze test dataset created by teachers," *arXiv:1711.03225*, 2017.

[36] W. Yu, Z. Jiang, Y. Dong, and J. Feng, "Reclor: a reading comprehension dataset requiring logical reasoning," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.

[37] J. Welbl, N. F. Liu, and M. Gardner, "Crowdsourcing multiple choice science questions," in *Proc. Workshop Noisy User Gener. Text*, 2017, pp. 94–106.

[38] S. K. Jauhar, P. D. Turney, and E. H. Hovy, "Tabmcq: A dataset of general knowledge tables and multiple-choice questions," *arXiv:1602.03960*, 2016.

[39] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 2381–2391.

[40] A. Horbach, I. Aldabe, M. Bexte, O. L. de Lacalle, and M. Maritxalar, "Linguistic appropriateness and pedagogic usefulness of reading comprehension questions," in *Proc. Int. Conf. Language Resour. Eval. (LREC)*, 2020, pp. 1753–1762.

[41] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv:1803.05457*, 2018.

[42] 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 *Proc. IEEE Conf. Comput. Vis Pattern Recognit. (CVPR)*, 2017, pp. 4999–5007.

[43] J. Berant, V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning, "Modeling biological processes for reading comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1499–1510.

[44] J. Shin, Q. Guo, and M. J. Gierl, "Multiple-choice item distractor development using topic modeling approaches," *Frontiers Psychol.*, vol. 10, p. 825, 2019.

[45] A. Trotman, A. Puurula, and B. Burgess, "Improvements to bm25 and language models examined," in *Proc. Australas. Document Comput. Symp. (ADCS)*, 2014.

[46] S. M. Downing, "Twelve steps for effective test development," *Handbook Test Develop.*, vol. 3, p. 25, 2006.

[47] S. Masapanta-Carrión and J. Á. Velázquez-Iturbide, "A systematic review of the use of bloom's taxonomy in computer science education," in *Proc. Spec. Interest Group Comput. Sci. Educ. (SIGCSE)*, 2018, pp. 441–446.

[48] S. M. Case and D. B. Swanson, *Constructing written test questions for the basic and clinical sciences*. NBME, 1998.

[49] L. Vanderwende, "The importance of being important: Question generation," in *Proc. Conf. Int. Natural Lang. Gener. (INLG)*, 2008.

[50] L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan, "Recent advances in neural question generation," *arXiv:1905.08949*, 2019.

[51] T. Tofade, J. Elsner, and S. T. Haines, "Best practice strategies for effective use of questions as a teaching tool," *Amer. J. Pharmaceutical Educ.*, vol. 77, no. 7, 2013.

[52] 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," *J. Mach. Learn. Res.*, vol. 21, no. 1, p. 5485–5551, 2020.

[53] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.

[54] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[55] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. Intrinsic Extrinsic Eval. Meas. Mach. Transl. Summarization*, 2005, pp. 65–72.

[56] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 8342–8360.

[57] A. Hadifar, S. Labat, V. Hoste, C. Develder, and T. Demeester, "A million tweets are worth a few points: Tuning transformers for customer service tasks," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2021, pp. 220–225.

[58] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2018.

[59] X. Sun, J. Liu, Y. Lyu, W. He, Y. Ma, and S. Wang, "Answer-focused and position-aware neural question generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 3930–3939.

[60] Z. Swiecki, H. Khosravi, G. Chen, R. Martinez-Maldonado, J. M. Lodge, S. Milligan, N. Selwyn, and D. Gašević, "Assessment in the age of artificial intelligence," *Comput. Educ.: Artif. Intell.*, vol. 3, p. 100075, 2022.

[61] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2017, pp. 785–794.

[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2019.

[63] S. Castro, R. Wang, P. Huang, I. Stewart, O. Ignat, N. Liu, J. Stroud, and R. Mihalcea, "Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2022, pp. 2925–2940.

[64] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[65] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. Autodiff Workshop*, 2017.

Amir Hadifar is a Ph.D. student at the Internet Technology and Data Science Lab (IDLab) at Ghent University. He is part of the Text-to-Knowledge (T2K) Group. His supervisors are Prof. Chris Develder and Prof. Thomas Demeester. He received his master's degree in Computer Engineering from the University of Amirkabir Tehran in 2018. His research interest includes AI in education and conversational systems.

Semere Kiros Bitew is a Ph.D. student at the Internet Technology and Data Science Lab (IDLab) at Ghent University. He is part of the Text-to-Knowledge (T2K) Group. His supervisors are Prof. Chris Develder and Prof. Thomas Demeester. He received his master's degree in Data Science & Smart services from the University of Twente, The Netherlands, in 2018. His research interest includes AI in education, controlled text generation, and psycho-linguistics.

Chris Develder is associate professor with the research group IDLab in the Dept. of Information Technology (INTEC) at Ghent Universityimec, Ghent, Belgium. He received the MSc degree in computer science engineering and a Ph.D. in electrical engineering from Ghent University (Ghent, Belgium), in Jul. 1999 and Dec. 2003 respectively (as a fellow of the Research Foundation, FWO). He has stayed as a research visitor at UC Davis, CA, USA (Jul.-Oct. 2007) and at Columbia University, NY, USA (Jan. 2013 - Jun. 2015). He was and is involved in various national and European research projects (e.g., FP7 Increase, FP7 C-DAX, H2020 CPN, H2020 Bright, H2020 BIGG, H2020 RENergetic, H2020 BD4NRG). Chris currently (co-)leads two research teams within IDLab, (i) UGent-T2K on converting text to knowledge (i.e., NLP, mostly information extraction using machine learning), and (ii) UGent-AI4E on artificial intelligence for energy applications (e.g., smart grid). He has co-authored over 200 refereed publications in international conferences and journals. He is Senior Member of IEEE, Senior Member of ACM, and Member of ACL.

Johannes Deleu received the Master of Science degree in computer science engineering from Ghent University, Belgium, in 2005. He is currently a Senior Research Engineer within the IDLab, Department of Information Technology, Ghent University-imec. His research concentrates on information extraction, machine learning, and in particular deep learning applied to natural language processing (NLP). He has participated in multiple research projects, developing automatic content enrichment systems for the media sector and more recently the education sector.

Thomas Demeester is an assistant professor at IDLab, at the Department of Information Technology, Ghent University-imec in Belgium. After his master's degree in electrical engineering (2005), he obtained his Ph.D. in computational electromagnetics, with a grant from the Research Foundation, Flanders (FWO-Vlaanderen) in 2009. His research interests then shifted to information retrieval (with a research stay at the University of Twente in The Netherlands, 2011), natural language processing (NLP) and machine learning (with a stay at University College London in the UK, 2016), and more recently to Neuro-Symbolic AI. He has been involved in a series of national and international projects in the area of NLP, and co-authored around one hundred peer-reviewed contributions in international journals and conferences. He is a member of AAAI and ELLIS.