Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy

Kevin Hwang*
Glenelg High School
Glenelg, MD, US 21737
kevinhwang7550@gmail.com

Sai Challagundla*
Glenelg High School
Glenelg, MD, US 21737
sschallag@gmail.com

Maryam M. Alomair
University of Maryland Baltimore County
Baltimore, MD, US 21250
maryama4@umbc.edu

Lujie Karen Chen
University of Maryland Baltimore County
Baltimore, MD, US 21250
lujiec@umbc.edu

Fow-Sen Choa
University of Maryland Baltimore County
Baltimore, MD, US 21250
choa@umbc.edu

Abstract

In educational assessment, Multiple Choice Questions (MCQs) are frequently used due to their efficiency in grading and providing feedback. However, manual MCQ generation encounters challenges. Relying on a limited set of questions may lead to item repetition, which could compromise the reliability of assessments and the security of the evaluation procedure, especially in high-stakes evaluations. This study explores an AI-driven approach to creating and evaluating MCQs in introductory chemistry and biology. The methodology involves generating Bloom's Taxonomy-aligned questions through zero-shot prompting with GPT-3.5, validating question alignment with Bloom's Taxonomy with RoBERTa–a language model grounded in transformer architecture, employing self-attention mechanisms to handle input sequences and produce context-aware representations of individual words within a given sentence–, evaluating question quality using Item Writing Flaws (IWF)–issues that can arise in the creation of test items or questions–, and validating questions using subject matter experts. Our research demonstrates GPT-3.5's capacity to produce higher-order thinking questions, particularly at the "evaluation" level. We observe alignment between GPT-generated questions and human-assessed complexity, albeit with occasional disparities. Question quality assessment reveals differences between human and machine evaluations, correlating inversely with Bloom's Taxonomy levels. These findings shed light on automated question generation and assessment, presenting the potential for advancements in AI-driven educational evaluation methods.

*Both Kevin Hwang and Sai Challagundla contributed equally.

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

1 Introduction

Multiple-choice questions (MCQs) have become a useful assessment tool in education [6, 5, 16]. Their effectiveness lies in their easy and efficient grading capacity, allowing educators to evaluate many responses efficiently. Furthermore, MCQs facilitate immediate feedback, which is invaluable for enhancing learning outcomes as it enables students to pinpoint areas of weakness and make timely improvements. Well-designed MCQs possess the remarkable ability to assess knowledge across different levels of Bloom's Taxonomy [1], a framework that classifies different levels of cognitive skills and abilities that students use to learn, thereby serving as a versatile tool to support and enhance learning outcomes. By aligning questions with different levels of Bloom's Taxonomy, instructors can control the cognitive depth of their questions, catering to the diverse learning needs of their students and encouraging critical thinking among learners [5].

The conventional method of generating and assessing questions has typically been demanding in terms of manual labor, often necessitating substantial human input and expertise. Furthermore, depending on a restricted pool of questions can result in the repetition of items, potentially undermining the reliability of assessments and the security of the assessment process. This constraint presents noteworthy difficulties, particularly in high-stakes assessment scenarios.

While automated question generation, particularly through the use of Large Language Models (LLM), presents a significant opportunity to streamline the question creation process, the potential to consistently generate high-quality MCQs aligned with Bloom's Taxonomy is an area that remains largely unexplored. The ability to harness the power of LLMs for precisely tailored MCQs that address different cognitive levels as defined by Bloom's Taxonomy offers a promising avenue for educational innovation. However, it also presents a set of unique challenges and complexities such as maintaining question quality, aligning with learning objectives, addressing biases, and ensuring scalability.

In this research, we investigate an AI-driven process for the creation and evaluation of MCQs in the domains of introductory chemistry and biology. This process comprises three components. First, we utilize zero-shot prompting to generate questions aligned with Bloom's Taxonomy, leveraging GPT-3.5, with a focus on contextual relevance within the disciplines. Second, we employ Natural Language Processing (NLP) techniques to evaluate the quality of these questions, assessing their alignment with Bloom's Taxonomy and adherence to Item Writing Flaw (IWF) [4] criteria, guidelines, and standards used in the field of educational assessment to evaluate the quality of test items or questions. Third, a subset of the questions was reviewed by a chemistry teacher with subject matter expertise and pedagogical insights. This validation procedure sought to bridge the gap between automated assessments and human standards, ensuring that the generated questions align with various taxonomic levels and could be utilized in classrooms.

This study seeks to address two research questions. RQ1 investigates the extent to which GPT-3.5 can generate questions that are aligned with the prompted Bloom's Taxonomy levels, as evaluated independently by both the machine learning model and a teacher. RQ2 explores how the quality of questions generated by GPT-3.5 compares when validated through expert judgment versus machine validation using the IWF criteria, and aims to discern the degree of alignment between these two validation methods. These research questions guide the inquiry into the efficacy and reliability of using advanced AI technologies for generating and validating high-quality MCQs in educational assessments.

2 Related Work

Previous studies in educational question generation have demonstrated the diverse applications of Pretrained Language Models (PLMs). [20] employed GPT-3, prompted with question-answer pairs sourced from an OpenStax biology textbook to generate MCQs and free-response questions (FRQs), as a notable example. They evaluated the quality of generated questions using the perplexity score (computed using GPT-2) and the grammatical error. In addition, they used subject-matter experts to determine that generated questions were ready to be used in classrooms. Another study by [17] involved fine-tuning Google T5 with questions extracted from an undergraduate data science course, evaluating question quality through information scores and GPT-3 classification. Meanwhile,[19] fine-tuned Google T5 on the Stanford Question Answering Dataset (SQuAD) to generate question-

answer pairs and used T5 and Sense2Vec to generate distractors, incorrect options meant to "distract" from the correct answer. However, they did not outline any attempts to evaluate the quality of generated questions. [2] expanded on this approach by fine-tuning Google T5 using a combination of datasets, including S2ORC, SQuAD, and SciQ. They evaluated the general questions for linguistic quality using the BLEU score and the F1 score and evaluated how human-like the questions were using the perplexity score and diversity score. They indicated that AI-generated questions could be easily repurposed by teachers for use in education. These approaches collectively highlight the versatility and adaptability of PLMs in generating educational questions, albeit without a specific focus on alignment with the distinct levels of Bloom's taxonomy. Notably, none of these methods have explored the potential of the widely accessible GPT-3.5 for this purpose.

In light of the existing research landscape, our study aims to address a notable gap in the literature concerning educational question generation. While previous research has demonstrated the utility of PLMs in generating educational questions, none of the mentioned methods have specifically targeted the alignment of questions with Bloom's taxonomy. In our comprehensive literature review, we have identified only two existing methods that have attempted to generate questions aligned with Bloom's taxonomy. [7] employs a template-based question generation system, using keyword identification and pattern matching to generate questions based on templates aligned with Bloom's taxonomy. [3] utilizes Bloom's taxonomy as a contextual template in prompt engineering for InstructGPT [10]. However, the first method does not validate Bloom's taxonomy post- generation, thereby lacking conclusive evidence regarding the ability of its approaches to generate questions that accurately correspond to Bloom's Taxonomy levels. The second method, though it does validate Bloom's Taxonomy, does not use GPT-3.5, generates free-response questions as opposed to MCQs, and does not evaluate questions using IWF criteria. Our paper introduces a novel contribution by investigating the capacity of GPT-3.5 to generate multiple choice questions that align specifically with Bloom's taxonomy, addressing this research gap and offering insights into the capabilities of NLP-based methods in educational question generation with a taxonomy focus.

3 Methods

3.1 Question Generation Strategy

We generated questions using OpenAI's GPT-3.5 model. To access GPT-3.5, we used the "GPT-3.5-turbo" model in the OpenAI Python library [11]. In our prompts, please refer to https://tinyurl.com/35am6sah for the template, we gave GPT-3.5 the section name for the selected questions, an excerpt about the topics, and a small excerpt describing each level of Bloom's taxonomy. Please refer to the supplementary details for our prompting template. We tasked GPT-3.5 with creating five MCQs with one correct answer and three distractors for each level of Bloom's taxonomy. For our question set, we generated questions based on college-level chemistry and biology textbooks (Chemistry 2e and Biology 2e from OpenStax [12, 13]. From each textbook, we extracted all the text from two randomly selected sections of chapters to use as excerpts for use in our prompts.

3.2 Question Quality Evaluation Strategy

We automatically evaluated the educational quality of generated questions with various NLP-based methods to detect common item writing flaws found in MCQs. Our detection system replicates [9]'s NLP-based IWF detector. The system detects a subset of 19 unique item-writing flaws based on 31 item-writing guidelines. If more than one item writing flaw is present, it is considered to be of low quality, as per [18].

3.3 Bloom's Taxonomy Evaluation Strategy

We use a machine learning model to classify the question's Bloom level to validate that GPT-3.5 is generating questions on the correct Bloom level. Our model and dataset are replicated from [14], however, we added callbacks to the model during training to control overfitting. The model's architecture consists of the RoBERTa LLM [8], a convolutional neural network, and a fully connected layer. Their dataset was sourced from four different papers and it consists of 2522 questions labeled with their Bloom level. We split the dataset into two stratified sets, one for training (90% of data) and one for testing (10% of data). During training, we implemented two callbacks: early stopping with a

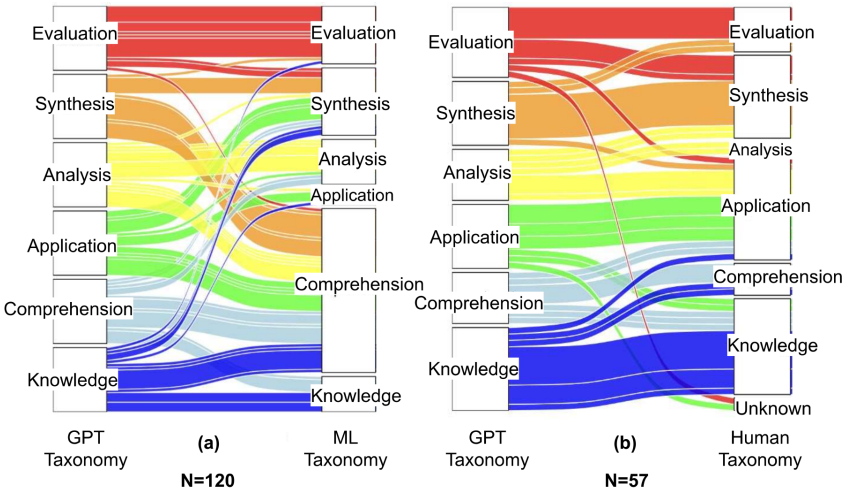

Figure 1: Shows the degree of alignment in Bloom's taxonomy between GPT-taxonomy and ML-taxonomy (subplot a with 120 samples) and between GPT-taxonomy and Human-taxonomy (subplot b with 57 samples). The width of the stripes indicates the number of matching questions at each Bloom's level, arranged from lowest ("knowledge") to highest ("evaluation").

patience of five epochs, and reducing the learning rate on a plateau with a patience of seven epochs. The model achieved an 83.79% accuracy and an 83.69% weighted F1 score on the testing set.

3.4 Human Evaluation

We utilized one domain expert with 28+ years of STEM teaching experience to evaluate the generated question set and validate our automated evaluation system. We gave the human rater a random subset of 57 out of the 120 questions generated. We provided the question, the correct answer, and the three distractors in the data. In addition, they were given all the excerpts that we used in generating the provided questions and the name of the chapter and section that each excerpt was sourced from. Using only the provided data, they were asked whether they would use the questions in a classroom setting (yes or no), the Bloom level of each question, and whether each question was relevant to its respective excerpt (yes or no).

4 Results

RQ1: To what extent does GPT-3.5 generate questions aligned with the prompted Bloom's Taxonomy levels, independently evaluated by the machine learning model and the human teacher? In this section, we present our analysis of Bloom's taxonomy levels of questions generated by GPT-3.5. We compared the levels GPT-3.5 was instructed to generate (GPT-taxonomy) with the levels predicted by machine learning models (ML-taxonomy) and the levels assigned by a teacher (Human-taxonomy) who was not informed of the use of GPT-3.5.

Figure 1 is a Sankey diagram [15] summarizing the degree of alignment among the GPT-taxonomy, ML-taxonomy, and Human-Taxonomy as defined above. Subplot (a) illustrates the alignment between the GPT-taxonomy and ML-taxonomy. As shown, 80% of questions at the "Evaluation" level per GPT-3.5 are classified at the same level by the machine learning model, while 70% of "Synthesis" questions are categorized as "comprehension." The remaining questions are classified at their respective levels. "Analysis" level questions are mostly divided between "Analysis" (50%) and "Comprehension" (40%) by the ML model. "Application" level questions are primarily categorized as "Synthesis" (35%) and "comprehension" (45%). For "Comprehension" questions, 50% are classified at the same level, whereas 25% are categorized as "Knowledge." Questions at the "Knowledge" level are split between "Knowledge" (30%) and "Comprehension" (45%) per ML model. Overall, most misalignments occur between adjacent levels. Per the ML model (i.e., ML-Taxonomy), the "comprehension" level

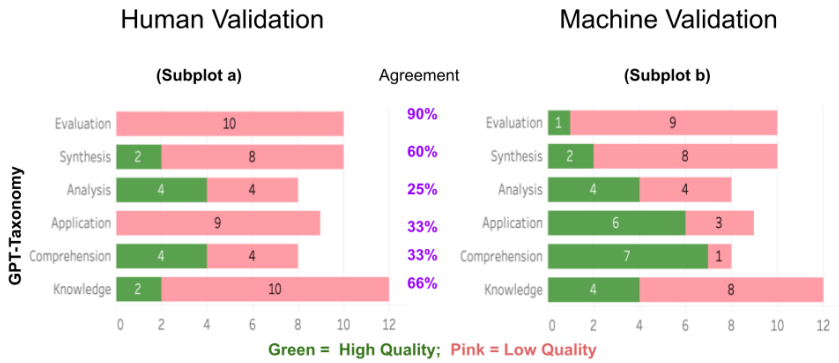

Figure 2: Shows the degree of alignment in Bloom's taxonomy between GPT-taxonomy and ML-taxonomy (subplot a with 120 samples) and between GPT-taxonomy and Human-taxonomy (subplot b with 57 samples). The width of the stripes indicates the number of matching questions at each Bloom's level, arranged from lowest ("knowledge") to highest ("evaluation").

dominates, which might be attributed to our relatively imbalanced training set where a majority of questions (38%) are at this level. Subplot (b) summarizes the alignment between GPT-taxonomy and Human-taxonomy. We note a high level of agreement, especially for the knowledge, application, and synthesis levels. Among the questions categorized in these levels by GPT-3.5, 75%, 78%, and 70%, respectively, were also determined to be at the same levels by the human teacher. When there was misalignment, the differences were mostly within 1 or 2 levels. For instance, about half of the questions categorized as "Evaluation" by GPT-3.5 were determined to be at the "synthesis" level by the human teacher. It's also worth noting that for a few questions categorized as "Application" and "Evaluation" by GPT-3.5, the human rater could not assign a specific level.

RQ2: How do the qualities of questions generated by GPT measure up when validated by a human teacher (Human validation) versus machine validation using the IWF criteria (Machine validation), and to what degree do these two validation methods align with each other? Figure 2 presents the quality evaluation results for a subset of 57 GPT-3.5-generated questions, assessed by the human teacher (left plot) and the machine (right plot) using the IWF criteria. The human teacher's rating is a binary indicator, determining if a given question is high quality and thus suitable for real-world usage. Meanwhile, the machine's evaluation relies on the IWF criteria, classifying a question as high quality if it meets at least 18 of the IWF criteria; otherwise, it is considered low quality. Based on the human validation results, only 12 out of the 57 questions are considered high-quality, with most coming from the comprehension and analysis levels (per GPT-Taxonomy). The IWF screening identified 24 (42%) questions as high quality, with a significant number from the application and comprehension levels. Except for the knowledge level, there is a general inverse relationship between Bloom's taxonomy level and quality. As one progresses up the taxonomy, GPT-3.5 finds it more challenging to produce high-quality questions. For instance, at the evaluation level, humans and machines found it unlikely to identify questions of high caliber. Regarding the agreement between human and machine evaluations, there's a notable variance, ranging from 25% (for analysis-level questions) to 90% (for evaluation-level questions).

5 Discussion

The analysis of Research Question 1 sheds light on the alignment of questions generated by GPT-3.5 with Bloom's Taxonomy levels, as assessed independently by a machine learning model (ML-taxonomy) and a teacher (Human-taxonomy). These findings offer valuable insights into the strengths and limitations of automated question generation in educational assessment.

5.1 Alignment Between GPT-taxonomy and ML-Taxonomy

The alignment analysis in RQ1 reveals GPT-3.5's potential to generate questions aligned with the intended Bloom's Taxonomy levels. Particularly promising is the strong alignment observed at the "evaluation" level, indicating GPT-3.5's capability to generate questions requiring higher-order thinking. However, some misalignments, especially between closely related levels such as "synthesis" and "comprehension," highlight the complexity of distinguishing between these cognitive levels. These misalignments underscore the need for continued refinement in GPT-3.5's question generation abilities.

5.2 Alignment Between GPT-taxonomy and Human-Taxonomy

The analysis between GPT-taxonomy and Human-taxonomy shows a notable level of agreement, particularly for the "knowledge," "application," and "synthesis" levels. This agreement implies that GPT-3.5 is generally proficient in generating questions that align with human perceptions of question complexity. However, occasional discrepancies, such as questions categorized as "evaluation" by GPT-3.5 being assessed as "synthesis" by the teacher, suggest room for improvement in distinguishing between higher-order cognitive levels. Additionally, some questions could not be assigned specific levels by the human teacher, indicating ambiguity in a subset of the generated questions.

5.3 Distribution of High-Quality Questions

The examination of Research Question 2 focuses on the assessment of the qualities of questions generated by GPT-3.5 through two distinct validation methods: human validation and machine validation using the IWF criteria. Additionally, this analysis explores the alignment between these two validation methods.

The results show that only a limited subset of the generated questions—12 out of 57—are considered high quality according to human validation. These high-quality questions are predominantly found within the comprehension and analysis levels, aligning with Bloom's Taxonomy as defined by GPT-3.5. In contrast, the machine's evaluation using IWF criteria identified a larger proportion—24 questions or 42%—as high quality, with an emphasis on questions from the application and comprehension levels. This discrepancy highlights differing perspectives on question quality between human judgment and the automated assessment process.

5.4 Cognitive Complexity and Question Quality

An interesting observation is the inverse relationship between Bloom's Taxonomy levels and question quality. As questions progress up the taxonomy, GPT-3.5 faces greater challenges in producing high-quality questions. For instance, at the evaluation level, both humans and machines found it less likely to identify questions of high caliber. This suggests that higher cognitive complexity levels, such as evaluation and synthesis, present greater difficulties for automated question generation.

5.5 Agreement Between Human and Machine Evaluations

The agreement between human and machine evaluations varies across different cognitive levels. Notably, there is a substantial level of agreement (90%) for evaluation-level questions, but this agreement diminishes as the cognitive complexity decreases. Analysis-level questions exhibit lower agreement (25%), indicating a higher level of variability in how human and machine assessors perceive their quality.

6 Conclusions

The results of RQ1 highlight GPT-3.5's potential as a tool for generating questions aligned with Bloom's Taxonomy levels. They underscore the importance of refining GPT-3.5's question generation capabilities, addressing nuances in cognitive level distinctions, and ensuring that automated assessments closely align with human standards and expectations in educational contexts. These findings pave the way for further exploration and enhancement of AI-driven question generation for educational assessment.

The quality evaluation results in RQ2 shed light on the challenges and nuances of automated question generation and quality assessment. The variation in assessments between human teachers and machine evaluators underscores the need for a balanced approach that considers both perspectives. The inverse relationship between cognitive complexity and question quality highlights the intricacies of generating high-quality questions at higher taxonomic levels. These findings have implications for improving the capabilities of automated systems in generating educational content, emphasizing the importance of refining automated question generation to align more closely with human standards and expectations.

In our future work, we will enhance question generation by using LangChain to automatically format GPT's responses. In addition, we will use few-shot learning with 5-shots, a method endorsed by Wang et al. We will upgrade to InstructGPT or GPT-4 for better alignment with human directions. To improve the automated validation of question quality, we will create an ML model using Item Writing Flaw (IWF) features to more closely mimic human verification. Our Bloom's taxonomy model will be further fine-tuned using GPT-generated questions, labeled using workers from Mechanical Turk. We'll assess questions using our current evaluation system, as well as linguistic quality through perplexity and diversity scores, and question relevance for specific contexts.

References

[1] Benjamin Bloom. A taxonomy of cognitive objectives. *New York: McKay*, 1956.

[2] Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer, 2023.

[3] Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

[4] Breakall Jared, Christopher Randles, and Roy Tasker. Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice 20*, 2:369–382, 2019.

[5] Arslaan Javaeed. Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. *MedEdPublish*, 7:128, 2018.

[6] Jindřich Klfa. Multiple choice question tests–advantages and disadvantages. *Recent Advances in Educational Technologies*, 2018.

[7] Selvia Ferdiana Kusuma and Rinanza Zulmy Alhamri. Generating indonesian question automatically based on bloom's taxonomy using template based method. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pages 145–152, 2018.

[8] Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer, 2021.

[9] Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning*, pages 229–245. Springer, 2023.

[10] Long OOuyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[11] OpenAI. openai-python: Openai's python library. `https://github.com/openai/openai-python`, 2020. Accessed: September 24, 2023.

[12] OpenStax. Openstax, 2019. Accessed: September 24, 2023.

[13] OpenStax. OpenStax, 2019. Accessed on September 24, 2023.

[14] Gani Mohammed Osman, Ayyasamy Ramesh Kumar, Sangodiah Anbuselvan, and Fui Yong Tien. Bloom's taxonomy-based exam question classification: The outcome of cnn and optimal pre-trained word embedding technique. *Education and Information Technologies*, pages 1–22, 2023.

[15] Ethan Otto, Eva Culakova, Sixu Meng, Zhihong Zhang, Huiwen Xu, Supriya Mohile, and Marie A Flannery. Overview of sankey flow diagrams: focusing on symptom trajectories in older adults with advanced cancer. *Journal of geriatric oncology*, 13(5):742–746, 2022.

[16] C Daniel Riggs, Sohee Kang, and Olivia Rennie. Positive impact of multiple-choice question authoring and regular quiz participation on student learning. *CBE—Life Sciences Education*, 19(2):ar16, 2020.

[17] Bhat Shravya, Huy Nguyen, Steven Moore, John Stamper, Majd Sakr, and Nyberg Eric. Towards automated generation and evaluation of questions in educational domains. In *the 15th International Conference on Educational Data Mining*, volume 701, Durham, UK, 2022.

[18] Marie Tarrant, Aimee Knierim, Sasha K Hayes, and James Ware. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26(8):662–671, 2006.

[19] Kristiyan Vachev, Momchil Hardalov, Georgi Karadzhov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. Leaf: Multiple-choice question generation. In *European Conference on Information Retrieval*, pages 321–328. Springer, 2022.

[20] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer, 2022.