



Revolutionizing Math Education: How Advanced Question Generation Streamlines Educator Workflows

Yongan Yu, Alexandre Krantz and Nikki G. Lobczowski

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 6, 2024

Revolutionizing Math Education: How Advanced Question Generation Empowers Students and Streamlines Educator Workflows

Anonymous submission

Abstract

Some educators have started to turn to Generative AI (GenAI) to help create new course content, but little is known about how they should do so. In this project, we investigated the first steps for optimizing content creation for advanced mathematics. In particular, we looked at the ability of GenAI to produce high-quality practice problems that are relevant to the course content. We conducted two studies to: (1) explore the capabilities of current versions of publicly available GenAI and (2) develop an improved framework to address the limitations we found. Our results showed that GenAI can create math problems at various levels of quality with minimal support, but that providing examples and relevant content results in better quality outputs. This research can help educators and institutions decide on the ideal way to adopt GenAI into their workflows, so it can be leveraged to create more effective educational experiences for students.

Introduction

With the rapid advancements in Large Language Models (LLMs), the intersection of artificial intelligence (AI) and educational development has emerged as an attractive global scholarly interest [38, 46, 28]. The versatile capabilities of LLMs have not only transformed traditional computational fields but have also significantly impacted the educational sector [2], upending how academic content is created, assessed and delivered [46]. Researchers such as Chen et al. (2020) [6], Denny et al. (2024) [8], and Li et al. (2024) [22] have demonstrated the potential of LLMs to enhance learning experiences and outcomes across various educational scenarios. In particular, recent studies have explored AI applications including cognitive diagnostics [9], knowledge tracking [36], personalized suggestions [19, 45], and learning behavior analysis [4, 30, 31].

While tools like ChatGPT have achieved widespread adoption with over 60% of educators experimenting with their use in classrooms, less than 20% feel adequately prepared to integrate these tools effectively into their curricula [17]. This significant discrepancy highlights a clear need for additional support to help educators fully leverage these advanced tools. One application of LLMs that could significantly support teachers is Question Generation (QG). A high-quality LLM QG tool could significantly reduce the workload of educators [7], as it would free up the time spent

on creating problem sets and answers. It could also result in more practice problems for students, enhancing their learning experience. From a technical standpoint, QG an existing sub-field of natural language processing (NLP), focused on enabling the automated creation of educational content directly from reference material, such as textbooks [10].

Despite the transformative potential of AI in education, its integration remains underutilized, especially in automated question generation. A need-finding study by Wang et al. (2023) [37] reveals that educators often express reservations about adopting AI tools, citing concerns about the relevance and quality of AI-generated content. Moreover, existing automatic QG tools are not widely used in classrooms due to their limited range in types and difficulty levels [20, 16]. Most systems primarily produce simple recall questions, failing to sufficiently challenge students or promote deeper cognitive processing.

Given these limitations and the potential for AI in education, our study aims to investigate the gap between AI capabilities and effective educational implementation. We also seek to answer (1) what impact increased contextual information has on the quality and cognitive depth of AI-generated questions and (2) how AI QG systems can be designed to provide outputs of varying cognitive depth. We hypothesize that by developing a context-aware AI framework integrating established educational taxonomies [18], we then can generate high-quality, diverse, and cognitively appropriate questions that align closely with specific educational objectives and content. Our goal is to bridge the gap between AI capabilities and educational needs, potentially increasing educator confidence and the adoption of AI tools in classrooms for improved educational outcomes.

Related-Work

Current Uses and Limitations of GenAI in Math Education

The rapid rise of GenAI, exemplified by ChatGPT's release in November 2022, has significantly impacted the educational landscape [24]. Within months, it became an integral tool for millions of students worldwide, prompting educational institutions to develop usage policies. By January 2023, it reached 100 million users [27] worldwide, making it the fastest-growing consumer application in history.

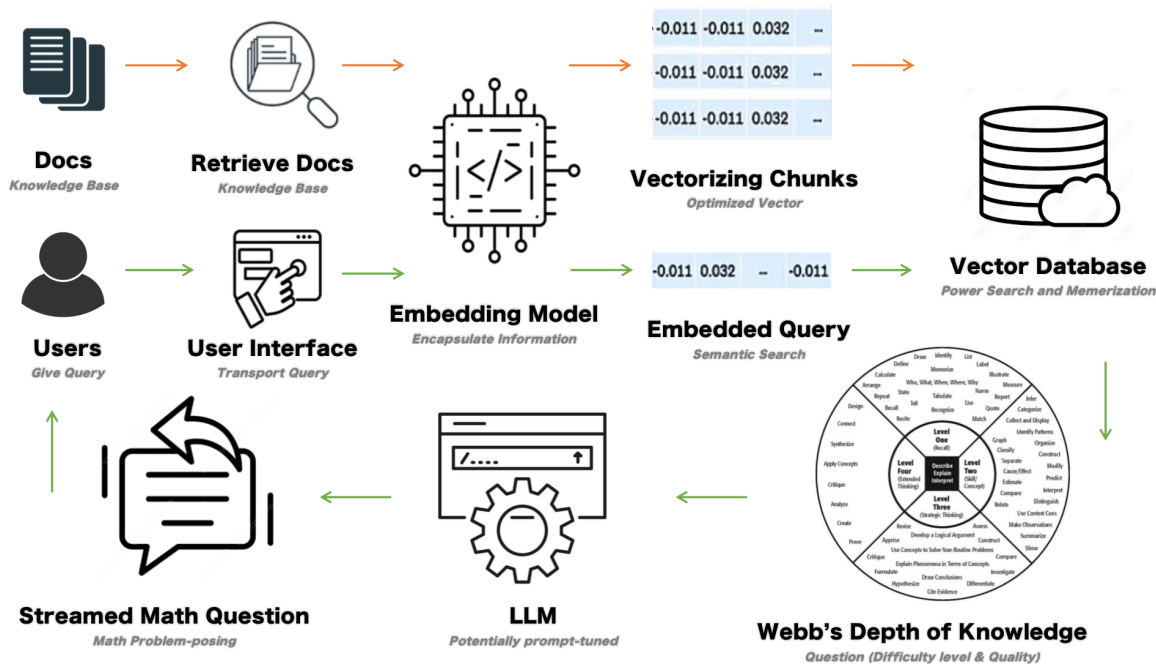


Figure 1: The Layout of DOK model

GenAI’s potential extends beyond education, with estimates suggesting it could accelerate general economic growth akin [25] to transformative technologies like electricity or the internet. In education specifically, it is projected that 20-40% of teachers’ administrative tasks could be partly or fully automated by AI [8], such as preparing learning materials, writing emails, and providing feedback to students [4]. However, current GenAI models have significant limitations [5, 33] in mathematics education. For example, while effective as a search engine for mathematical theorems and structures, ChatGPT struggles as a conversational tutor [13] for complex or obscure math problems. Frieder et al. (2023) also found that even with elementary-level problems, these models can make errors [13] such as ignoring the order of operations or making misleading conceptual statements.

Despite surface-level familiarity with mathematical topics, these models often lack the rigid logical reasoning [26] required for advanced mathematics. These limitations highlight the need for more specialized and context-aware AI systems in mathematics education [20], particularly in areas like automated question generation.

Evolution of Question Generation (QG) Models through LLMs

A variety of QG systems have been developed, often for educational purposes, e.g., assisting students in reading, vocabulary learning, or the assessment of reading comprehension [7]. Initially, QG systems primarily focused on some simple recall questions [39], which limited their utility within the field of education. These early models utilized basic

rule-based algorithms [32] to generate problems from static content with little advantage in terms of adaptability or depth. With the rapid development of machine learning and deep learning techniques, models have begun to utilize NLP to parse and interpret educational material, allowing for a broader range of question types beyond simple recall. For instance, Sutskever et al. (2014) [35] developed a reader-generator framework that enhances the adaptability of QG systems to different educational contents, showing a 30% improvement in question relevance compared to traditional models. This framework enabled the encoding process to integrate not just the textual information but also the targeted answer into its context, thereby improving the relevance of question generation.

Recently, a data-driven QG model was developed by Wang et al. (2018) [40] called “QG-Net.” QG-Net is specifically tailored for generating quiz questions from educational content, leveraging a recurrent neural network (RNN) architecture to automate this process. Our research primarily started with this model for question generation, but we soon realized that this QG-Net model greatly relied on extensive training data, a challenge exacerbated by the typically small size of benchmark datasets [16] in QG, leading to overfitting issues and redundancy in a set of questions [15]. Specifically, it is even harder to retrieve the data since we investigate math problem-posing, as there are more benchmarks. For instance, while general language tasks might have benchmarks with millions of examples, mathematics-specific QG datasets might only contain a few thousand carefully curated problems [44], spanning various topics and

difficulty levels.

Thus, by pursuing this direction, we not only aim to improve the immediate quality and relevance of generated questions but also to lay the groundwork for future adaptive learning environments where AI can dynamically respond to individual student progress and needs.

Purpose of Current Study

This research centers on two main objectives using a sequential study research design: 1) Explore the capabilities of current versions of publicly available GenAI, and 2) Develop an improved framework to address the limitations found. To address these aims, we used an exploratory approach to investigate the quality of content by varying the parameters of the input and manually evaluating the output, based on certain criteria.

Study 1

For the first study of our research, we conducted an initial exploration to test the capabilities of GenAI in creating accurate and relevant mathematics questions corresponding to different types of questions. As part of the investigation, we varied the level of contextual information provided to the AI, with the aim of determining the optimal amount of additional context. This initial exploration served as a foundation for understanding the potential and limitations of GenAI in educational content creation.

Exploring Current Capabilities of GenAI

We began by exploring how GenAI can support current educational practices, such as Intelligent Tutoring Systems (ITS) and regular quizzing [34]. The aim was to identify an application that can seamlessly integrate into the workflow of students and teachers and functions with high levels of reliability and context awareness. Thus, we decided to look at whether GenAI could create comprehension questions that are relevant to the course content, and examine the cognitive depth of those comprehension questions using Bloom's Taxonomy [12]. This test aims to assess how varying levels of contextual information affect the quality, relevance, and cognitive depth of AI-generated questions.

To this day, Bloom's Taxonomy remains one of the primary models for categorizing various levels of understanding in educational contexts. The 2001 revision of Bloom's Taxonomy [1] presents a hierarchical structure of cognitive processes, with each level building upon the previous ones. The six levels, from lowest to highest, are: Remember, Understand, Apply, Analyze, Evaluate and Create. In a typical learning process, learners progress from one level to the next, as each level is a prerequisite for the next. In mathematics education, this progression is particularly relevant as students move from memorizing formulas to applying them in novel situations and eventually creating new mathematical proofs or models. Therefore, we chose Bloom's Taxonomy here because it is widely recognized and used in educational settings, making our results more interpretable and applicable for educators. It also provides distinct levels in a

straightforward framework and corresponds well with cognitive processes involved in learning and applying mathematical concepts.

The hierarchical nature of Bloom's Taxonomy for categorizing understanding also has its critics, though. In particular, the hierarchical nature of the stages is contested. It is said to be an oversimplification [11] of the holistic and interconnected nature of learning. However, this criticism can be applied to all models of mental processes. We used Bloom's Taxonomy really as a tool to categorize the depth of the questions. By allowing the model to generate questions from any level of the hierarchy, we are not constricting the user from following Bloom's hierarchy in their learning process.

Study Design The independent variable here is *the level of context* provided to the GenAI. We tested three scenarios that instructors could engage in, each one adding additional context to the prior. We chose a course about *Mathematical Logic*, as the name implies, that requires extensive and rigid logical reasoning, the topics cover the following topics: *satisfiability*, *compactness theorem*. We then designed prompts that applied the model to act as a teaching assistant to create a set of five comprehension questions with answers, that are relevant to the course content and help students review the material, maintained the same prompts for three scenarios. We also the Gemini 1.5 Pro model [29] for each scenario, chosen for its large context window and ability to process substantial input text. To ensure consistency, we maintained a low-temperature setting of 0.2 across all scenarios [47], minimizing random variations in output. The hypothesis here was that additional context would increase the output quality, but we were unsure about how strong the effects would be.

- **Scenario 1 (Minimal Context):** the teacher will upload the syllabus of the class and indicate which topic they have taught in the class in 1-2 sentences. Our model will generate content based on its understanding of the topic, from the initial training data educator provided in this case.
- **Scenario 2 (Moderate Context):** the teacher will upload the syllabus and their notes for a specific class. Our model will have some content to analyze, but will not know the full content of everything that has been covered in the class. Still, we hypothesize that the questions will be more relevant to the material covered in class.
- **Scenario 3 (Comprehensive Context):** the teacher will upload the syllabus, their notes from a particular class, and a reference text that covers all the material that is covered by the class, throughout the semester. The expectation here is that our model will have a greater understanding of the context of the lesson and be able to use the relevant terminology and reference previous topics.

Evaluation Metrics The output scenario sample is shown in Figure 2. To assess the quality of the generated questions, we incorporated an expert review process to evaluate the correctness of the GenAI outputs. A mathematics student with

Question: Can a theory have arbitrarily large finite models but no infinite model?

Answer: No, if a theory has arbitrarily large finite models, the Compactness Theorem guarantees the existence of an infinite model.

Question: Give an example of a theory that is finitely satisfiable but not satisfiable.

Answer: There is no such theory; the Compactness Theorem proves that finite satisfiability implies satisfiability.

Figure 2: Two of the five outputs for Scenario 2.

| Context Scenario | Relevance | Depth | Correctness |
|------------------|--------------------|--------------------|--------------------|
| Minimal | 1.00 ± 0.00 | 2.60 ± 1.14 | 0.60 ± 0.55 |
| Moderate | 1.00 ± 0.00 | 2.60 ± 1.82 | 0.80 ± 0.45 |
| Comprehensive | 0.80 ± 0.45 | 2.40 ± 2.07 | 1.00 ± 0.00 |

Table 1: Performances comparisons for three context scenarios, the best-performing scoring implementation is in bold for each metric. Note: Values are presented as mean ± standard deviation.

expertise in Mathematical Logic reviewed and developed a scoring system based on three key metrics:

- **Relevance:** A binary score (0 or 1) indicating whether the question was within the scope of the lesson.
- **Depth:** A score from 0 to 6, corresponding to the levels of Bloom’s Taxonomy, with higher scores indicating deeper cognitive engagement.
- **Correctness:** A binary score (0 or 1) assessing the accuracy of the provided answer.

Analysis Methods We calculated average scores for each metric across the five questions generated in each scenario and presented them as mean ± standard deviation. The result is shown in Table 1, which allowed us to compare the performance of the GenAI model under different levels of contextual information.

Findings from Initial GenAI Testing

Our initial exploration focused on testing GenAI’s capability to generate mathematics questions with answers aligned with Bloom’s Taxonomy under varying levels of contextual information. In this preliminary experimentation, a few sample outputs are shown in Figure 2, we found that GenAI indeed can create content that is relevant and high-quality with input support. We then evaluated the performances through the methods that are seen before, the results are summarized in Table 1. Results from testing were surprising in the level of relevance because there was a declining trend in the relevance level of the output as we added additional context. However, we found the depth of the questions remained relatively constant, and the correctness of the answers generated improved as more context was provided, reaching perfect accuracy with comprehensive context.

Advancing to the Next Stage The principal finding from this investigation is that the additional context was able to improve the correctness but not necessarily the relevance of the output. Although the AI is informing itself from the context provided, it is also extrapolating and going beyond the bounds of what was taught in the course (i.e., creating “hallucinations”). Given that most instructors are likely able to calculate answers to the generated questions, we decided to focus more on relevance than correctness moving forward. We considered using a different technical architecture, such as Retrieval-Augmented Generation (RAG) [21], which could potentially resolve this issue by forcing the AI to source its generation from a “chunk” (or section) of the materials provided. Additionally, the depth of the generated questions varied, which led to an average between levels 2 and 3. However, there was no user control over what level of question was generated. The natural next step would be to investigate a design where GenAI creates questions aimed at a certain level, on Bloom’s Taxonomy or some other learning framework.

DOK VS. Bloom’s Taxonomy In addressing the challenge of generating high-quality math questions, researchers have previously explored various frameworks to categorize cognitive demands. Bloom’s Taxonomy, widely recognized and used across disciplines to classify educational goals and objectives, provided a crucial starting point [41]. However, while Bloom’s Taxonomy offers a comprehensive structure for understanding cognitive processes, we found that it may not fully capture the nuanced cognitive demands specific to mathematics education.

The DOK framework, developed by Norman Webb (2002), offered several advantages for mathematical education. Primarily, DOK emphasizes the complexity of tasks rather than just the type of thinking, aligning better with mathematical problem-solving. It takes into account the context in which knowledge [41] is applied, which is crucial for mathematical applications. DOK’s four levels - recall and reproduction, skills and concepts, strategic thinking, and extended thinking - provide a more nuanced approach to assessing the depth of understanding required in mathematics.

Furthermore, DOK’s alignment with curriculum standards and its widespread use in assessment design make it particularly relevant to our goal of creating an AI system that can generate questions closely aligned with educational objectives. This framework allows for a more precise mapping of question difficulty to the cognitive processes involved in mathematical reasoning, from basic recall to complex problem-solving.

Study 2

Building upon prior findings, we implemented an advanced framework integrating RAG with Webb’s Depth Of Knowledge [42]. This approach aims to address the limitations identified in our initial phase and enhances the quality and relevance of AI-generated educational content in mathematics.

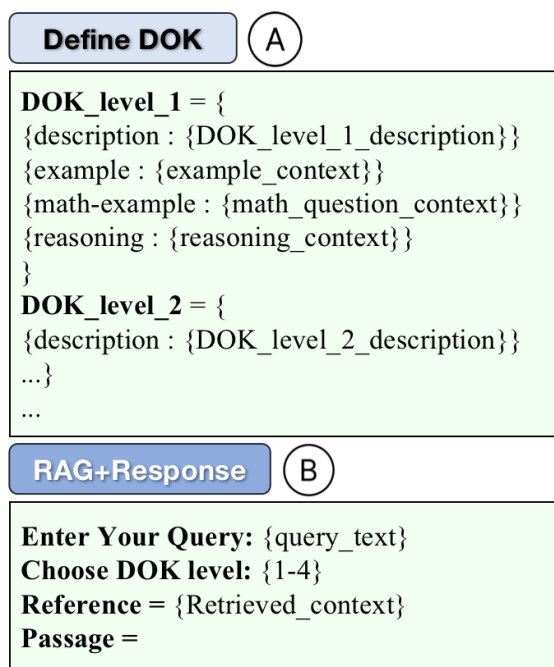


Figure 3: (A) shows a level prompt template example, due to space constraints; the complete template covers all four DOK levels. (B) represents the user input interface, with outputs displayed in the 'Passage' section.

Developing an Improved Framework

Our second research goal focuses on enhancing the utility and applicability of LLMs for educational purposes, particularly by developing a QG model that leverages DOK levels. We explored how QG models can generate different difficulty levels of questions from a given query that teachers and students find useful in mathematics education. This phase of our research not only outlines our rationale for this approach but also describes the implementation of our improved system. The results of this study aim to foster greater integration of AI in educational contexts and broader applications in general education enhancements.

Retrieval-Augmented Generation (RAG) Framework

To address the limitations identified in our initial exploration, particularly the lack of contextual relevance, we implemented a *Naive RAG* framework [14]. This approach consists of three key components: Retrieval, Augmentation, and Generation:

- **Retrieval:** We created a vector database of mathematical content, including materials, textbooks, and practice examples (see orange line in Figure 1). When a query is input, the system searches through our database of mathematical content to find the most relevant information. This step utilizes a vector embedding model of both the query and the stored content, allowing for semantic similarity matching rather than just keyword matching.
- **Augmentation:** The retrieved information is then used to enhance the context provided to the language model. This

Recall and Reproduction

Enter your query: Intermediate Value Theorem
Enter the Depth of Knowledge (DOK) level (1-4): 1

In a math tutorial, a student was studying the Intermediate Value Theorem. He came across the equation $x = e(x - 2)$ and was asked to find a solution in the interval $(0, 1)$. He also had to split this interval into two halves $(0, 1/2)$ and $(1/2, 1)$ and determine which of these two intervals has a solution to the equation. he student also studied the equations $\cos x = x$ on $(0, 1)$ and $\ln x = e^{-x}$ on $(1, 2)$. Based on the above information, answer the following questions:

1. What is the Intermediate Value Theorem?
2. What does it mean to find a solution in a given interval?
3. What is the purpose of splitting the interval into two halves?
4. What does the equation $\cos x = x$ represent?
5. What does the equation $\ln x = e^{-x}$ represent?

Figure 4: The test sample in DOK level 1

crucial step ensures that the model has access to specific, relevant mathematical knowledge when generating questions, significantly reducing the likelihood of producing irrelevant or incorrect content.

- **Generation:** The augmented context is fed into our pre-trained large language model to generate appropriate questions at the desired DOK level. This process allows for the creation of questions that are not only relevant to the specific mathematical topic but also aligned with the intended cognitive depth (see green line in Figure 1).

Integration of DOK levels We incorporated four DOK levels [42] into our question generation process, as illustrated in Figure (A) in Figure 3. Each level was defined and prompted with a description, relevant contexts, practice examples, and rationale for how questions fit the level [42]. The descriptions of the four levels are:

- **Recall and Reproduction (DOK Level 1):** This level involves recalling information such as facts, definitions, terms, or simple procedures. It also includes performing basic algorithms or applying straightforward formulas. Key verbs associated with this level include "identify," "recall," "recognize," "use," and "measure".
- **Skills and Concepts (DOK Level 2):** At this level, students are required to make decisions about how to approach a problem. Unlike Level 1, which involves rote responses or following set procedures, Level 2 demands more cognitive engagement. Key verbs include "classify," "organize," "estimate," and "make observations".
- **Strategic Thinking (DOK Level 3):** This level requires students to use reasoning, planning, and evidence. It involves a higher level of thinking than the previous two levels. Activities that require students to make conjectures typically fall into this category.

| # COG | Relevance | | | Depth Accuracy | | | Appropriateness | | |
|---------|-----------|-------------|--------------|----------------|-------------|-------------|-----------------|--------|--------------|
| | ZS | FS+DOK | FS+DOK+R | ZS | FS+DOK | FS+DOK+R | ZS | FW+DOK | FW+DOK+R |
| Level 1 | 0.72 | 0.85 | 0.84 | 0.59 | 0.81 | 0.85 | 0.92 | 0.80 | 0.91 |
| Level 2 | 0.84 | 0.81 | 0.85 | 0.79 | 0.74 | 0.30 | 0.81 | 0.72 | 0.90 |
| Level 3 | 0.66 | 0.79 | 0.81 | 0.71 | 0.72 | 0.81 | 0.62 | 0.74 | 0.95 |
| Level 4 | 0.57 | 0.68 | 0.80 | 0.60 | 0.69 | 0.82 | 0.68 | 0.70 | 0.82 |
| Average | 0.6975 | 0.7825 | 0.825 | 0.6725 | 0.74 | 0.6925 | 0.7575 | 0.74 | 0.895 |

Table 2: Performance of different structure testing

- **Extended Thinking (DOK Level 4):** The highest level requires students to make multiple connections—relating ideas within the content area or across content areas. Students must select from various approaches to solve complex problems, demonstrating the highest level of cognitive demand.

Figure 4 provides an example output for a DOK Level 1 question related to the *Intermediate Value Theorem* in Calculus. This sample illustrates how our system generates questions that align with the recall and reproduction level, focusing on basic definitions and concepts without requiring deeper analysis or application.

Implementation Details We implemented the core framework in *Python*, serving the *GPT-4* model for question generation with default temperature settings, accessed via the OpenAI API. Following Figure 1, we preprocessed and embedded a corpus of math content using the *text-embedding-ada-002* model from OpenAI. Documents were broken into fixed-size chunks with sliding windows to improve semantic relevance. As such, we used *Pinecone*, hosted on *AWS*, to store vectors and facilitate advanced search functionalities. The user interface of our system for now is designed for simplicity and effectiveness. As shown (B) in Figure 3, users are prompted to input two key pieces of information, the mathematical intuition or concept they want to explore, and the desired DOK level, ranging from one to four. Based on the retrieval strategies, our QG-DOK system generates tailored questions, seamlessly integrating the specified cognitive depth with the relevant mathematical content.

Evaluation Metrics To assess the performance of our QG-DOK framework, we conducted a comparative evaluation involving math students. We evaluated our method by comparing model performance on a held-out test set [43] across three different implementations. We established a zero-shot baseline, where the rubric was included in the prompt without any labeled examples. Next, we implemented a few-shot baseline, which provided the model with labeled instances to illustrate how DOK functions, as depicted in Figure 3. Our final baseline incorporated both few-shot learning and the RAG framework. As table 2 illustrated, the incremental baselines were as follows: Zero-Shot (ZS), Few-Shot with DOK (FS+DOK), and Few-Shot with DOK

and RAG (FS+DOK+R).

To compare implementations, we selected metrics similar to those used in our initial test for creating high-quality math questions. We conducted evaluations with GPT-4 using G-Eval [23], designed to align closely with human judgment. According to instructions of G-Eval, the evaluation was designed to focus on three binary metrics, each scored on a scale from zero to one. The metrics chosen were 1) relevance score which indicates whether the generated question was related to the initial query, 2) DOK alignment with cognitive levels, contributing to the quality of the math question and 3) appropriateness whether the question is suitable for students at their current level, considering the context of the query. Importantly, we decided not to prompt or evaluate answers to the questions given the poor results in the previous stage and the likelihood that instructors could determine suitable answers themselves.

Findings from Improved Work Overall, our evaluation of the QG-DOK framework demonstrates promising results across key metrics, indicating significant improvements in question generation for math education. As shown in Table 4, relevance scores show a clear upward trend as we move from the ZS baseline to our full FS+DOK+R model. The average relevance increased from 0.6975 in ZS to 0.7825 in FS+DOK, and further to 0.825 in FS+DOK+R. This 18.3% improvement from ZS to FS+DOK+R indicates that our framework significantly enhances the contextual alignment of generated questions with the initial query. Notably, the relevance scores for higher DOK levels 3 and 4 saw the most substantial improvements, with level 4 questions increasing from 0.57 in ZS to 0.80 in FS+DOK+R, a 40.4% boost. Depth Accuracy results present a more nuanced picture. While the overall average improved from 0.6725 in ZS to 0.74 in FS+DOK, it slightly decreased to 0.6925 in FS+DOK+R. This unexpected result warrants further investigation and refinement, particularly for level 2 questions where accuracy dropped significantly from 0.74 in FS+DOK to 0.30 in FS+DOK+R. However, levels 3 and 4 saw considerable improvements in the FS+DOK+R model. Appropriateness shows the most consistent improvement across all levels. The average score increased from 0.7575 in ZS to 0.895 in FS+DOK+R, with an 18.2% improvement. This metric saw enhancements across all DOK levels, with partic-

ularly notable gains in Levels 3 and 4. For instance, Level 3 questions improved from 0.62 in ZS to 0.95 in FS+DOK+R, a 53.2% increase.

Our analysis of the output quality revealed some persistent challenges, particularly in the accuracy of the representation of mathematical equations. A notable example is evident in Figure 4, where the equation $x = e(x-2)$ appears instead of the correct form $x = e^{(x-2)}$. This error stems from a limitation in our data preprocessing: the source PDF document lacked the crucial exponential symbol $\hat{}$, leading to its omission during the RAG process. This error highlights a critical consideration for future implementations. LaTeX formatting, which is widely used in mathematical typesetting, appears to be more suitable for LLMs in question generation tasks. We found LaTeX's explicit markup of mathematical symbols and structures could prevent such misinterpretations, ensuring that complex equations are accurately represented and retrieved.

Otherwise, the FS+DOK+R model consistently outperformed other implementations in relevance and appropriateness, demonstrating the value of integrating both DOK frameworks and RAG techniques. However, the mixed results in depth accuracy, particularly at Level 2, indicate an area for further refinement in our model. Despite this, the overall performance suggests that our QG-DOK framework represents a significant advancement in automated question generation for mathematical education, particularly for higher-order thinking skills.

Future Research Pathways We introduced the QG-DOK framework integrating Webb's DOK prompting with the RAG technique and demonstrated significant improvements over our initial exploration. The framework addressed key limitations identified in Study 1, notably enhancing question relevance (18.3% improvement), cognitive depth alignment (particularly for higher DOK levels), and pedagogical appropriateness (18.2% increase). These advancements indicate substantial progress in generating mathematically relevant, cognitively appropriate, and educationally suitable questions, particularly for promoting deeper cognitive processing.

Despite these achievements, areas for further refinement are still required. The unexpected decrease in depth accuracy for Level 2 questions and the slight overall decline in depth accuracy when implementing RAG highlight the need for better consistency across DOK levels. Additionally, while the framework shows an interesting pattern in reducing hallucinations and improving contextual awareness, further work is needed to enhance the evaluation of mathematical content quality and to assess the framework's scalability across various mathematical topics. In summary, Study 2 represents a significant step forward in AI-assisted question generation for mathematics education, while also illuminating paths for future research and development.

Discussion

Our research, which was conducted through two interconnected studies, provides significant insights into the potential and challenges of using GenAI for mathematics educa-

tion, particularly in question generation. Study 1 explored the capabilities of GenAI in creating mathematics questions and answers aligned with Bloom's Taxonomy under varying levels of contextual information. This initial exploration revealed that while GenAI showed promise in generating relevant questions, it struggled with consistently producing questions at higher cognitive levels.

Overall, the clarity and quality of the GenAI comprehension questions were very encouraging. Regardless of the context provided, though, the LLMs struggled to consistently ask cognitively advanced questions, which led to instabilities in the depth of the questions generated. In Table 1, we observed that as more context was provided in the comprehensive scenario, the model increasingly generated plausible yet incorrect information, indicating that LLMs are prone to creating hallucinations.

Building on these findings, Study 2 introduced the QG-DOK framework, integrating Webb's DOK levels with a RAG approach. We found that by leveraging resources that educators are already familiar with, both the quality and depth of the generated questions can be improved [3]. Additionally, our findings support earlier studies suggesting that AI can effectively generate educational content but requires careful design to ensure cognitive depth and relevance [8, 22]. The improvements seen in Study 2 address limitations identified in previous research, particularly in generating deeper thinking questions.

Limitations and Future Work Although our results demonstrated the potential of GenAI in educational content creation, a few limitations occurred that must be addressed. First, the hierarchical nature of Bloom's Taxonomy for categorizing understanding has its critics, specifically concerning the oversimplification of the holistic and interconnected nature of learning [11], which mirrored our concerns. Switching to DOK to inform the AI for question generation helped alleviate this to a degree to meet our exploratory goals, but it is still possible that simply providing a description of each level is not enough instruction for the AI to create an appropriate question. An interesting avenue for future research could be to investigate problem design in greater depth. In particular, one could break down what makes a good problem by studying how teachers go about designing problem sets. Then, one could chart out a step-by-step process for the AI to follow when generating problems.

Finally, we acknowledge the constraints of our data input and evaluation methods. Specifically, we used a singular reference text for each study, with streamlined human and AI evaluations. Given the exploratory nature of this study, it was important to limit the parameters to best explain variations in the quality. Moving forward, researchers can test and evaluate the output more systematically, given that our findings have highlighted the capabilities (and limitations) of current genAI. As such, our work serves as an important preliminary step in advancing question generation through AI for advanced mathematics.

Acknowledgement

[Blinded for Review]

References

- [1] Anderson, L. W.; and Krathwohl, D. R. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.
- [2] Antu, S. A.; Chen, H.; and Richards, C. K. 2023. Using LLM (Large Language Model) to Improve Efficiency in Literature Review for Undergraduate Research. *LLM@ AIED*, 8–16.
- [3] Ayanwale, M. A.; Sanusi, I. T.; Adelana, O. P.; Aruleba, K. D.; and Oyelere, S. S. 2022. Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, 3: 100099.
- [4] Batool, S.; Rashid, J.; Nisar, M. W.; Kim, J.; Kwon, H.-Y.; and Hussain, A. 2023. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1): 905–971.
- [5] bin Mohamed, M. Z.; Hidayat, R.; binti Suhaizi, N. N.; bin Mahmud, M. K. H.; binti Baharuddin, S. N.; et al. 2022. Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education*, 17(3): em0694.
- [6] Chen, L.; Chen, P.; and Lin, Z. 2020. Artificial intelligence in education: A review. *Ieee Access*, 8: 75264–75278.
- [7] De Kuthy, K.; Kannan, M.; Ponnusamy, H. S.; and Meurers, D. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5786–5798.
- [8] Denny, P.; Prather, J.; Becker, B. A.; Finnie-Ansley, J.; Hellas, A.; Leinonen, J.; Luxton-Reilly, A.; Reeves, B. N.; Santos, E. A.; and Sarsa, S. 2024. Computing education in the era of generative AI. *Communications of the ACM*, 67(2): 56–67.
- [9] Dias, R.; and Torkamani, A. 2019. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, 11(1): 70.
- [10] Elkins, S.; Kochmar, E.; Cheung, J. C.; and Serban, I. 2024. How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23084–23091.
- [11] Fadul, J. A. 2009. Collective Learning: Applying Distributed Cognition for Collective Intelligence. *International Journal of Learning*, 16(4).
- [12] Forehand, M. 2010. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4): 47–56.
- [13] Frieder, S.; Pinchetti, L.; Griffiths, R.-R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.; and Berner, J. 2024. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36.
- [14] Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- [15] Guo, S.; Liao, L.; Li, C.; and Chua, T.-S. 2024. A survey on neural question generation: Methods, applications, and prospects. *arXiv preprint arXiv:2402.18267*.
- [16] Heilman, M. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.
- [17] Hojeij, Z.; Kuhail, M. A.; and ElSayary, A. 2024. Investigating in-service teachers' views on ChatGPT integration. *Interactive Technology and Smart Education*.
- [18] Irvine, J. 2021. Taxonomies in education: Overview, comparison, and future directions. *Journal of Education and Development*, 5(2): 1.
- [19] Kumar, V.; Rajan, B.; Venkatesan, R.; and Lecinski, J. 2019. Understanding the role of artificial intelligence in personalized engagement marketing. *California management review*, 61(4): 135–155.
- [20] Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; and Al-Emari, S. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30: 121–204.
- [21] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- [22] Li, H.; Xu, T.; Zhang, C.; Chen, E.; Liang, J.; Fan, X.; Li, H.; Tang, J.; and Wen, Q. 2024. Bringing generative AI to adaptive learning in education. *arXiv preprint arXiv:2402.14601*.
- [23] Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- [24] Lo, C. K. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4): 410.
- [25] McAfee, A. 2024. Generally Faster: The Economic Impact of Generative AI.
- [26] McCarthy, J. 2022. Artificial intelligence, logic, and formalising common sense. *Machine Learning and the City: Applications in Architecture and Urban Design*, 69–90.
- [27] Milmo, D. 2023. ChatGPT reaches 100 million users two months after launch— Chatbots— The Guardian. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>.
- [28] Pedro, F.; Subosa, M.; Rivas, A.; and Valverde, P. 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development.

- [29] Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillcrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- [30] Romero, C.; and Ventura, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1): 135–146.
- [31] Romero, C.; and Ventura, S. 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, 40(6): 601–618.
- [32] Sarker, I. H. 2022. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2): 158.
- [33] Shin, D. 2020. An analysis prospective mathematics teachers’ perception on the use of artificial intelligence (AI) in mathematics education. *Communications of Mathematical Education*, 34(3): 215–234.
- [34] Shute, V. J.; and Psotka, J. 1994. 19. INTELLIGENT TUTORING SYSTEMS: PAST, PRESENT, AND FUTURE.
- [35] Sutskever, I. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv preprint arXiv:1409.3215*.
- [36] Tsui, E. 2003. Tracking the role and evolution of commercial knowledge management software. *Handbook on Knowledge Management: Knowledge Directions*, 5–27.
- [37] Wang, D.; Tao, Y.; and Chen, G. 2024. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123: 102275.
- [38] Wang, N.; and Johnson, M. ??? AI Education for K-12: Connecting AI Concepts to High School Math Curriculum. *Workshop on Education in Artificial Intelligence K-12, 28th International Joint Conference on Artificial Intelligence*.
- [39] Wang, X.; Fan, S.; Houghton, J.; and Wang, L. 2022. Towards process-oriented, modular, and versatile question generation that meets educational needs. *arXiv preprint arXiv:2205.00355*.
- [40] Wang, Z.; Lan, A. S.; Nie, W.; Waters, A. E.; Grimaldi, P. J.; and Baraniuk, R. G. 2018. QG-net: a data-driven question generation model for educational content. In *Proceedings of the fifth annual ACM conference on learning at scale*, 1–10.
- [41] Weay, A. L.; Masood, M.; and Abdullah, S. 2016. Systematic review of revised Bloom Taxonomy, SOLO Taxonomy and Webb’s Depth of Knowledge (DOK) in assessing students’ historical understanding in learning history. *Malaysian Journal of Higher Order Thinking Skills in Education*, 3: 1–27.
- [42] Webb, N. L. 2002. Depth-of-knowledge levels for four content areas. *Language Arts*, 28(March): 1–9.
- [43] Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- [44] Wu, H.; Hui, W.; Chen, Y.; Wu, W.; Tu, K.; and Zhou, Y. 2023. Conic10K: a challenging math problem understanding and reasoning dataset. *arXiv preprint arXiv:2311.05113*.
- [45] Xiong, Z.; Li, H.; Liu, Z.; Chen, Z.; Zhou, H.; Rong, W.; and Ouyang, Y. 2024. A review of data mining in personalized education: Current trends and future prospects. *Frontiers of Digital Education*, 1(1): 26–50.
- [46] Zhai, X.; Chu, X.; Chai, C. S.; Jong, M. S. Y.; Istenic, A.; Spector, M.; Liu, J.-B.; Yuan, J.; and Li, Y. 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021(1): 8812542.
- [47] Zhu, Y.; Li, J.; Li, G.; Zhao, Y.; Jin, Z.; and Mei, H. 2024. Hot or Cold? Adaptive Temperature Sampling for Code Generation with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 437–445.