

Math Multiple Choice Question Generation via Human-Large Language Model Collaboration

Jaewook Lee¹, Digory Smith², Simon Woodhead², Andrew Lan¹
University of Massachusetts Amherst¹, Eedi²
{jaewooklee, andrewlan}@cs.umass.edu
{digory.smith, simon.woodhead}@eedi.co.uk

ABSTRACT

Multiple choice questions (MCQs) are a popular method for evaluating students' knowledge due to their efficiency in administration and grading. Crafting high-quality math MCQs is a labor-intensive process that requires educators to formulate precise stems and plausible distractors. Recent advances in large language models (LLMs) have sparked interest in automating MCQ creation, but challenges persist in ensuring mathematical accuracy and addressing student errors. This paper introduces a prototype tool designed to facilitate collaboration between LLMs and educators for streamlining the math MCQ generation process. We conduct a pilot study involving math educators to investigate how the tool can help them simplify the process of crafting high-quality math MCQs. We found that while LLMs can generate well-formulated question stems, their ability to generate distractors that capture common student errors and misconceptions is limited. Nevertheless, a human-AI collaboration has the potential to enhance the efficiency and effectiveness of MCQ generation.

Keywords

Multiple Choice Question, Large Language Models, Human-in-the-loop.

1. INTRODUCTION

Multiple choice questions (MCQs) are widely used to evaluate students' knowledge since they enable quick and accurate administration and grading [2, 6, 9]. MCQs are constructed in a specific format. The *stem* refers to the statement on the problem setup and context, followed by a question that needs to be answered. Among the options, the correct one can be referred to as the *key*, while incorrect ones can be referred to as *distractors*. As the name implies, distractors in MCQs are typically formulated to align with common errors among students. These distractors are chosen because students either i) lack the necessary comprehension of the *knowledge components (KCs)* or concepts/skills tested in the

question to accurately identify the key as the correct answer or ii) exhibit misconceptions that make them think a specific distractor is correct.

While MCQs offer many advantages in student knowledge assessment, manually crafting high-quality MCQs, especially in math-related domains, is a demanding and labor-intensive process [5]. There are three main tasks in this process: First, educators need to formulate a question stem that effectively encapsulates the KCs they aim to test. Second, educators need to anticipate common errors and/or misconceptions among students and create corresponding distractors. Third, educators need to provide feedback to students who select distractors that can help them identify their errors and lead them to the correct answer, to expedite their learning process.

The emergence of large language models (LLMs) has raised hopes for making MCQ creation more scalable by automating the process. Specifically, few-shot, in-context learning is promising for generating math MCQs since LLMs can follow instructions based on contextual information conveyed by a few examples. While automated question generation for open-ended questions has shown notable success, generating plausible distractors within MCQs presents a different challenge: distractors should be based on anticipated student errors/misconceptions [12], whereas LLMs have not necessarily learned this information during training. Moreover, math MCQs are challenging since they require mathematical reasoning, which means that distractors cannot be generated using a knowledge graph [13] or paraphrasing tool [8]. Consequently, math educators need to take an important role in guiding LLMs in math MCQ generation: LLMs are responsible for scaling up the process while humans use their expertise efficiently. Therefore, we raise following are two core research questions (RQs) that help identify opportunities to generate math MCQs through collaboration between LLMs and human educators: 1) RQ1: Can LLMs generate valid MCQs, especially distractors and feedback corresponding to common student errors/misconceptions? 2) RQ2: What are the key design elements in a system where human math educators and LLMs collaborate on MCQ generation?

1.1 Contributions

In this paper, we introduce a prototype tool called the Human Enhanced Distractor Generation Engine (HEDGE) for math MCQ creation, which leverages the expertise of educators by asking them to edit LLM-generated MCQs in a two-step

J. Lee, D. Smith, S. Woodhead, and A. Lan. Math multiple choice question generation via human-large language model collaboration. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 941–946, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zenodo.12730005>

