The “Task” at Hand: LLMs and Low-Floor, High-Ceiling Math Tasks

Anna Gustaveson and Margaret Ann Donnelly (University of North Carolina)

Introduction

The National Council for Teachers of Mathematics (NCTM) outlines practices for conceptually-oriented and equitable math teaching (Leinwand, et. al., 2012). Teachers may implement these practices through low-floor, high-ceiling math tasks that engage students in a common math problem using strategies, tools, and representations of their choice. An example math task in Figure 1 includes both a low floor where students can solve the problem in multiple ways (e.g., using fraction bars or multiplication) and a high ceiling with multiple solution paths and possible answers. These tasks promote teaching practices outlined by NCTM such as supporting reasoning and problem-solving, building procedural fluency from conceptual understanding, and centering students’ ideas.

Figure 1

Mystery Fractions

Sam added 3 fractions together for a total of $\frac{5}{4}$. Each fraction had a different number in the denominator, and each number in the denominator was less than 10. What are the possible fractions Sam used to make $\frac{5}{4}$?

However, finding or crafting quality math tasks can be challenging. Advancements in generative AI tools like Large Language Models (LLMs) have been shown to support task creation (Küchemann, et. al., 2023), but less is known about the teaching practices LLM-created tasks promote. If math teachers use LLMs for planning, it is important that Teacher Educators (TEs) understand the types of tasks LLMs generate and the practices they support. This study asks: *To what extent do two popular LLMs construct low-floor, high-ceiling tasks that promote NCTM’s productive practices?*

Methods

ChatGPT-3.5 (GPT) and Gemini, both accessible, open-source LLMs, were prompted to: “Create a low floor, high ceiling math task about adding fractions with unlike denominators.” Both LLMs generated a scenario with related mathematical prompts for students to solve (see Figure 2) as well as teacher notes about how to teach the task.

Figure 2

Low-floor, high-ceiling math tasks generated by LLMs

ChatGPT	Gemini
Task Description: You are planning a party and need to make fruit punch. You have two different recipes for fruit punch, each requiring different fractions of fruit juices. Your task is to create your own fruit punch recipe by adding fractions with unlike denominators. **Low Floor Version:** Imagine you have a recipe that calls for 1/4 cup of orange juice and another recipe that calls for 1/3 cup of pineapple juice. Your task is to add these two fractions together to figure out how much fruit juice you need for your punch. Use visual aids like fraction circles or fraction bars if needed. **High Ceiling Extensions*** 1. Recipe Variation: Introduce additional fruit juice fractions into the mix, such as 1/5 cup cranberry juice or 1/6 cup grape juice. Challenge students to add multiple fractions with unlike denominators to create their own unique fruit punch recipe. 2. Fraction Conversion: Present the task using mixed numbers or improper fractions instead of proper fractions. For example, instead of ¼ cup orange juice, use 2/4 or 1/2 cup of orange juice. This adds complexity and encourages students to convert between different forms of fractions. *High Ceiling Extensions 3-5 not included for space	**Scenario:** It's pizza party time! You and your friends order two different pizzas: - A large pizza cut into eight slices (each slice is 1/8 of the pizza) - A medium pizza cut into slices (each slice is 1/6 of the pizza) There are leftover slices: - 3 slices from the large pizza - 2 slices from the medium pizza **Challenge:** 1. How much pizza is leftover altogether? (Low Floor) 2. Can you represent the leftover pizza using fractions? (Low Floor) 3. What if you wanted to combine all the leftover slices on one plate? How much pizza would that be? (Higher Ceiling) 4. Can you express the total leftover pizza as a fraction with a denominator of your choice? (High Ceiling) 5. (Bonus) Can you think of other ways to represent the total leftover pizza? (e.g., decimals, percentages) (Very High Ceiling)

We first solved the tasks and used analytic memoing to note structural elements, strategy potential, mathematical accuracy, and language use (Figure 3). In a second round of coding, we highlighted portions of the tasks and teacher notes that did or did not align with NCTM's productive teaching practices marking each task as "yes" for alignment, "no" for misalignment, or "n/a" for no evidence of alignment or misalignment (Leinwand, et. al., p. 24, Figure 4).

Figure 3

Analytic Memo Summaries

	Chat-GPT 3.5	Gemini
Summary of notes about math task structure	The low floor and high ceiling in the task are separate versions rather than part of one problem every student can access	In the task, there are five math questions for students, marked with the level of question rather than one problem with entry points built in
Summary of notes about strategies	The task encourages use of visual tools but doesn't specify what strategy students should use; Multiple entry points/strategies are possible	The teacher notes describe what strategy students could use for each of the leveled questions (e.g. question 3 encourages use of converting to common denominator of 24); Multiple entry points limited by mathematical inaccuracy
Summary of notes about mathematical accuracy	There is a mathematical error in the task for High Ceiling Extension 4 (i.e., when it says $\frac{1}{4} + \frac{3}{12} = \frac{1}{3} + \frac{1}{2} = 1\ cup$)	The fractional amounts in the task refer to different sized wholes, so the fractions can't be added If wholes were the same, there is still a mathematical error in the teacher notes (i.e., when it says $\frac{3}{8} + \frac{16}{24} = \frac{15}{24}$)
Summary of notes about language use	Misuse of "mixed number" and "improper fraction" in High Ceiling Extension 2	Misuse of "mixed number" in the teacher notes Not enough differentiation in language across leveled questions to know how to respond

Figure 4

Alignment with NCTM Practices

	NCTM Productive Teaching Practices							
	Establishing Math Goals to Focus Learning	Implement Tasks that Promote Reasoning and Problem Solving	Use and Connect Mathematical Representations	Facilitate Meaningful Mathematical Discourse	Pose Purposeful Questions	Build Procedural Fluency from Conceptual Understanding	Support Productive Struggle in Learning Mathematics	Elicit and Use Evidence of Student Thinking
GPT	Yes	Yes	Yes	N/A	N/A	Yes	N/A	N/A
Gemini	No	No	N/A	N/A	No	No	No	N/A

Findings

ChatGPT-3.5

GPT's task was mathematically accurate and aligned with several of the productive teaching practices. In a teacher note, the task encouraged varied approaches, stating, "*Encourage students to use multiple strategies… Emphasize the importance of explaining their reasoning and justifying their answers.*" The high ceiling extensions posed open-ended problems without a suggested math strategy and directed teachers to ask students to explore ideas without taking over their thinking. However, GPT's math task was not without its shortcomings. It interpreted low floor and high ceiling as separate problems. The low-floor prompt suggested specific strategies and numbers, while the high-ceiling prompts allowed for a range of both. Splitting the task could offer students differential access to rich problems depending on the version assigned. Additionally, linguistic errors (i.e., misusing improper fractions) and mathematical errors (i.e., incorrectly adding fractions together) were both present in the task.

Gemini

Gemini's task was neither mathematically accurate nor did it align with productive teaching practices. The task's teacher notes stated which representations and tools they should have students use for each question (e.g. *"find a common denominator (24)"*), limiting student reasoning. Additionally, the task had mathematical and language errors. To add fractional pieces together, they must refer to the same sized whole, yet in the pizza task, students are asked to add together the leftover slices from two different-sized pizzas. Though a mathematically rich conversation could be had about whether you can add fractions together from different sized wholes, a teacher simply following the task might assume that the slices can be combined. Another error occurs in the teacher notes: "*Advanced students can convert the final fraction [15/24] to a mixed number (5/8).*" As 15/24 is not a fraction greater than 1, its equivalent 5/8 is not going to be a mixed number. More likely a language error than a mathematical one, Gemini has substituted the language of "converting to a mixed number" for "finding a simplified equivalent fraction". This is not the only place where Gemini's language is erroneous. The model asks four leveled questions about leftover pizza. Even when assuming that the word-problem pizzas were same-sized wholes and attempted questions 1-3 as written can be answered with the same responses.

Discussion

In this study we asked two LLMs to generate a low-floor, high-ceiling math task for adding fractions, and the implications can be summed up simply: if elementary math teachers choose to use LLMs to generate complex tasks for students, they must engage in effective prompting and evaluating the LMM's output before classroom implementation.

TEs should consider supporting teachers with prompt engineering to get the most out of LLMs and emphasize the importance of testing AI-suggested tasks to confirm they are mathematically sound. While "low-floor, high-ceiling" tasks engage students in the same problem through multiple entry points, the LLMs read the prompt as either two separate task versions for students of varying abilities or as a series of questions purportedly differentiated as low-floor or high-ceiling. More careful prompt engineering may be needed (e.g., giving the LLMs a more detailed explanation of a low-floor, high-ceiling task) to generate better responses.

Additionally, there were language and math errors from both LLMs, and they made Gemini's task unusable for the intended mathematical goal of the lesson. Database information being pulled in to support task creation can contain mathematical errors (e.g. trying to add different sized wholes) and linguistic errors (e.g. mixed number in place of equivalent fraction), so prompt engineering may not be as useful in avoiding these errors. Instead, teachers must be prepared to notice errors in AI-generated materials before presenting them to students.

In addition to prompting and evaluating AI-generated tasks, teachers must also be prepared to implement them well. Previous scholarly research indicates that while AI can converse in signs and signals, it is unable to communicate the human *experience* of 'doing the thing' (i.e., enacting a math task, Frick, 2024). Therefore, we must carefully consider how LLMs direct teachers to use tasks. GPT encouraged the use of multiple strategies and representations, aligning with several of NCTM's productive practices. In contrast, Gemini directed teachers to ask funneling questions based on specific procedures. However, enactment of either task is up to the teacher. It is crucial that TEs continue to build teachers' pedagogical expertise so they may translate AI-generated directions, questions, and concepts into teaching and learning practice in productive ways.

It might be simpler to conclude that TEs should not encourage teachers to use LLMs in making math tasks for their classrooms, but as AI is becoming more ubiquitous in nature, we argue that it is worth learning from these early models and preparing teachers for possible uses. The potential of using LLMs to support math task creation is still largely untapped, but TEs can support teachers use of these tools through intentional prompting, evaluation, and implementation of LLM-created math tasks.

References

Frick, T. W. (2024). Are we dupes? Limitations of AI systems: What should educators do with them? *TechTrends: Linking Research & Practice to Improve Learning*, *68*(1), 14–26. https://doi.org/10.1007/s11528-023-00893-3

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development?. *Physical Review Physics Education Research*, *19*(2), 020128. http://dx.doi.org/10.1103/PhysRevPhysEducRes.19.020128

Leinwand, Steve. *Principles to actions: Ensuring mathematical success for all*. National Council of Teachers of Mathematics, Incorporated, 2014.