Counterexamples to Demonstrate Artificial Intelligence Chatbot's Lack of Knowledge in the Mathematics Education Classroom

Amanda Gantt Sawyer and Zareen Gul Aga (James Madison University)

Artificial Intelligence (AI) is a part of our society and has become normalized in our college classrooms. However, researchers have identified how AI chatbot responses can be biased and inaccurate (Marr, 2023; Wu, 2023). Teachers are still using AI chatbots (e.g., ChatGPT) as curriculum creation tools, making AI the newest mathematics education curriculum developer (Gerstenschlager & Marin, 2024; Sawyer, 2024). In investigating elementary mathematics preservice teachers' (PSTs) use of ChatGPT in creating mathematics tasks, Sawyer (2024) found preservice teachers were overconfident in the AI chatbot's abilities. PSTs accepted AI-generated tasks as perfect and made little to no adaptations in mathematics content or pedagogy. One PST explained, "It's like a calculator for mathematics resources" (Sawyer, 2024, p. 21). Since PSTs may misunderstand what the AI chatbot can accomplish, Mathematics Teacher Educators (MTE) might consider breaking PSTs' trust in the tool by showcasing the AI chatbot's failures. Thus, in this paper, we share counterexamples that MTEs can showcase demonstrating ChatGPT's inaccurate mathematics information and that ChatGPT's text responses are not always appropriate for students.

ChatGPT's Inaccurate Mathematical Information

ChatGPT is a Large Language Model (LLM) trained to generate responses by predicting the next word in a series using text from its database (OpenAI, 2023). Thus, ChatGPT 3.5 does not think arithmetically and needs help with simple calculations (Marcus, 2023). As seen in the first counterexample, ChatGPT was prompted to multiply 953 by 987, which is 940,611, yet the LLM concluded that the value was 941,961. The example demonstrates how AI chatbots will present invalid mathematics as facts, and PSTs must be wary of any mathematics it creates. The program is designed to provide unique answers and follows steps from its training rather than mathematical logic; thus, it can provide invalid mathematical responses.

Figure 1: ChatGPT's Inaccurate Mathematics

Prompt to ChatGPT: What is 953*987?

ChatGPT's Response: 953 multiplied by 987 equals 941961.

In addition, ChatGPT may not recognize specific mathematical education concepts. For example, analyzing curriculum materials for their level of cognitive demand (Stein & Smith, 1998) or categorizing PSTs' mathematical questioning (Boaler & Broadie, 2004) are examples of topics taught in many mathematics methods or content courses. PSTs need to be made aware that ChatGPT will provide an answer to questions about those topics without understanding the meaning of those terms. When we asked how it defined the level of cognitive demand, ChatGPT did not use Stein and Smith's (1998) Task Analysis Guide Framework but created an understanding using

Bloom’s Taxonomy. As seen in the second counterexample, Figure 2, ChatGPT responded based on lower to higher-order thinking skills. We caution teachers not to make assumptions about the universal use of mathematics education terms. Likewise, MTEs could specify to their PSTs that specific terminology must be defined to the LLM when requesting mathematics education-related prompts to get a more accurate response. It would benefit the users of ChatGPT to be careful when asking questions about mathematical work and checking the AI chatbot’s definitions.

Figure 2: ChatGPT’s Lack of Mathematical Education Concepts

Prompt to ChatGPT: What is the level of cognitive demand of the task you created?

ChatGPT’s Response: The cognitive demand of the task I created is moderate. The task involves multiple steps and requires students to perform basic multiplication operations. Students need to understand the context of the problem, identify relevant information, and apply multiplication to find the solutions. The cognitive demand is increased by incorporating additional elements, such as the introduction of a pattern (every 3rd plot) and the calculation of seeds produced by sunflowers.

Overall, the task encourages students to engage in analytical thinking and apply mathematical concepts in a real-world context. It falls within the range of moderate cognitive demand as it involves several sequential steps, but it doesn't require advanced problem-solving strategies or extensive reasoning beyond basic multiplication skills.

Next Prompt to ChatGPT: What are the levels of cognitive demand?

ChatGPT’s Response: The levels of cognitive demand refer to a framework used in education to describe the cognitive complexity or difficulty of tasks that students are asked to perform. The concept is often associated with Bloom's Taxonomy, which classifies cognitive skills into a hierarchical structure. The original Bloom's Taxonomy has six levels, arranged from lower-order thinking skills to higher-order thinking skills:

1. **Remembering:** Involves recalling facts, information, or procedures.

2. **Understanding:** Involves demonstrating comprehension and understanding of concepts and ideas.

3. **Applying:** Involves using acquired knowledge and understanding in new or different situations.

4. **Analyzing:** Involves breaking down information into parts and examining the relationships between them.

5. **Evaluating:** Involves making judgments based on criteria and standards.

6. **Creating:** Involves generating new ideas, products, or ways of viewing things.

Educators use these levels to design instructional activities and assessments that align with the intended cognitive demands of a lesson or course. The goal is to gradually move students from lower-order thinking skills to higher-order thinking skills as they progress in their learning.

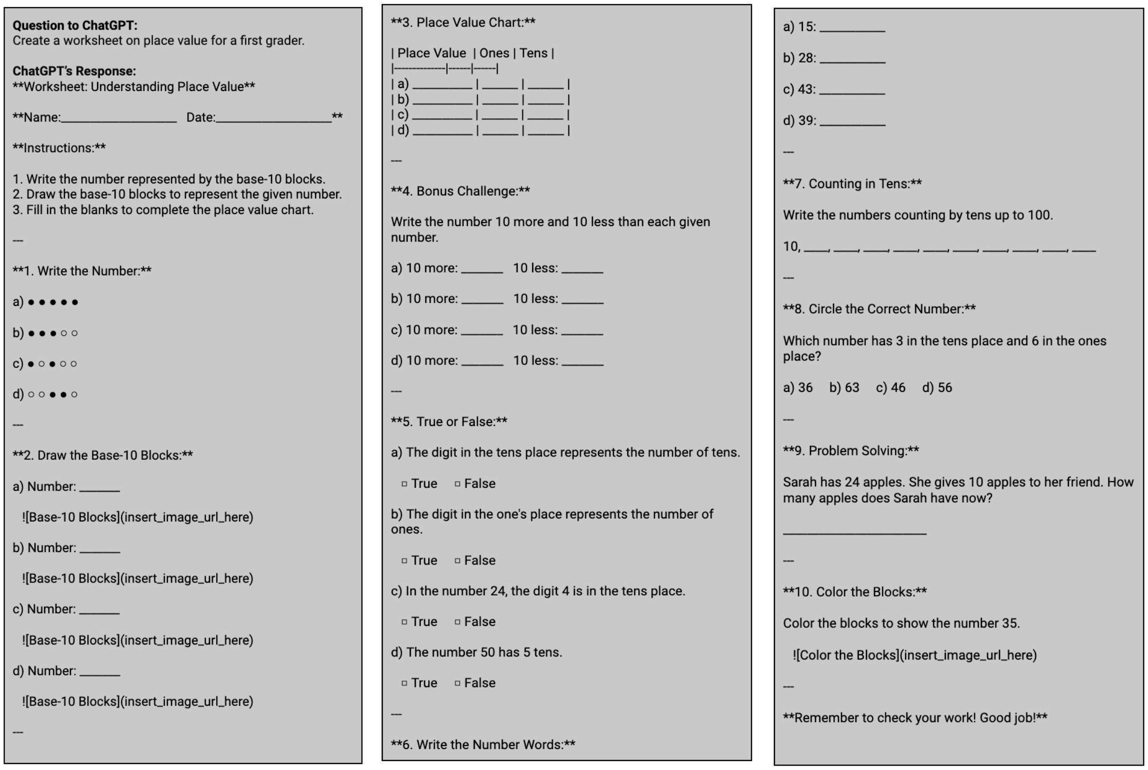
It's important to note that there are variations and adaptations of Bloom's Taxonomy, and some models include additional levels or use different terminology to describe cognitive demand. The levels provide a useful framework for educators to create learning objectives and assessments that promote deeper understanding and critical thinking skills.

ChatGPT's Inappropriate Materials for Students

ChatGPT's responses were designed for users over 18 years of age (OpenAI, 2023). Thus, ChatGPT has difficulty creating language-appropriate worksheets for younger students. As seen in Figure 3, the worksheet ideas generated by the LLM cannot be directly implemented in a classroom. Many of the activities do not make sense, for example, in the third counterexample:

> Write the number 10 more and 10 less than each given number.
>
> a) 10 more: _______ 10 less: _______

Figure 3: ChatGPT's Developmentally Inappropriate Materials



Question to ChatGPT:
Create a worksheet on place value for a first grader.

ChatGPT's Response:

```
**Worksheet: Understanding Place Value**

**Name:________________   Date:________________**

**Instructions:**

1. Write the number represented by the base-10 blocks.
2. Draw the base-10 blocks to represent the given number.
3. Fill in the blanks to complete the place value chart.

---

**1. Write the Number:**

a) ● ● ● ● ●

b) ● ● ● ○ ○

c) ● ○ ● ○ ○

d) ○ ○ ● ● ○

---

**2. Draw the Base-10 Blocks:**

a) Number: ______

  ![Base-10 Blocks](insert_image_url_here)

b) Number: ______

  ![Base-10 Blocks](insert_image_url_here)

c) Number: ______

  ![Base-10 Blocks](insert_image_url_here)

d) Number: ______

  ![Base-10 Blocks](insert_image_url_here)

---

**3. Place Value Chart:**

| Place Value  | Ones | Tens |
|--------------|------|------|
| a) _________ | _____ | _____ |
| b) _________ | _____ | _____ |
| c) _________ | _____ | _____ |
| d) _________ | _____ | _____ |

---

**4. Bonus Challenge:**

Write the number 10 more and 10 less than each given number.

a) 10 more: ______   10 less: ______

b) 10 more: ______   10 less: ______

c) 10 more: ______   10 less: ______

d) 10 more: ______   10 less: ______

---

**5. True or False:**

a) The digit in the tens place represents the number of tens.

  ▫ True   ▫ False

b) The digit in the one's place represents the number of ones.

  ▫ True   ▫ False

c) In the number 24, the digit 4 is in the tens place.

  ▫ True   ▫ False

d) The number 50 has 5 tens.

  ▫ True   ▫ False

---

**6. Write the Number Words:**

a) 15: _________

b) 28: _________

c) 43: _________

d) 39: _________

---

**7. Counting in Tens:**

Write the numbers counting by tens up to 100.

10, ____, ____, ____, ____, ____, ____, ____, ____, ____, ____

---

**8. Circle the Correct Number:**

Which number has 3 in the tens place and 6 in the ones place?

a) 36   b) 63   c) 46   d) 56

---

**9. Problem Solving:**

Sarah has 24 apples. She gives 10 apples to her friend. How many apples does Sarah have now?

____________________

---

**10. Color the Blocks:**

Color the blocks to show the number 35.

  ![Color the Blocks](insert_image_url_here)

---

**Remember to check your work! Good job!**
```

ChatGPT did not provide a "given number" to create the values, making the activity meaningless. In addition, asking a first-grade student to complete a three-page worksheet on a concept may not be developmentally appropriate (Ransom & Manning, 2013). We recommend that PSTs be cautioned not to implement any worksheets created by the AI without critically curating the activity and adapting it for their students. In particular, elementary PSTs could benefit from learning ways to ensure that the AI-generated content is safe and developmentally appropriate for their students.

We found some of the AI-generated content to be inappropriate not just for elementary students but also for middle school students. For example, ChatGPT generated a seventh-grade mathematics lesson on how to teach proportional relationships to solve multistep ratio and percent problems. ChatGPT's lesson related proportions to real-world scenarios with step-by-step instructions for implementing the activity. However, as seen in the fourth counterexample, Figure 4, the "real-life scenario card" is inappropriate for seventh-grade students. Notice that the scenario described a weight loss program for middle school students to understand proportional reasoning. Researchers identified that promoting weight loss progress in schools can harm students' physical and emotional health (Hanson, 2023); thus, teachers should not present this example in their classes. This example demonstrates how AI chatbots currently do not have human discretion to choose appropriate concepts for different grade levels. While ChatGPT can construct lesson plans quickly that can relate mathematical ideas to real-world scenarios, PSTs must be aware that it is their responsibility to ensure that the AI's lack of knowledge does not harm their students.

Figure 4 : ChatGPT's Inappropriate Content

Question to ChatGPT: Create the real-life scenario card for students described in your lesson.

ChatGPT's response: **Real-Life Scenario Card: Weight Loss Progress**

Scenario: A person is on a weight loss program and has lost 10 pounds in 2 months. If they continue to lose weight at this rate, how much weight will they have lost in 6 months?

These scenario cards can be distributed to students in pairs or small groups for them to work on during the lesson. Encourage students to use the concept of proportional relationships to solve these real-life problems and present their solutions to the class.

Be Cautious with AI

PSTs will use these AI chatbots both currently and in their future classrooms. MTEs can guide their PSTs to be cautious because of these tools' limitations. AI chatbots are not calculators for mathematics resources, and PSTs must be made aware of this fact. MTEs might consider teaching their PSTs to thoughtfully select mathematics resources based on their pedagogical knowledge, content knowledge, personal experiences, and the lesson's purpose (Sawyer et al., 2020). Therefore, consider using these counterexamples to support PST's understanding of the capabilities of these exciting new tools.

References

Association of Mathematics Teacher Educators. (2017). 'Standards for preparing teachers of mathematics.' *Association of Mathematics Teacher Educators*, amte.net/standards, page accessed 1.10.2024.

Boaler, J., & Brodie, K. (2004, October). The importance, nature, and impact of teacher questions. In *Proceedings of the twenty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 774–782).

Hanson, O. (2023). 'Healthy eating' curriculum can do more harm than good. *CNN Health.* https://www.cnn.com/2023/10/09/health/unhealthy-school-nutrition-lessons-wellness/index.html

Gerstenschlager, N. E., & Marin, K. A. (2024). GPS: Supporting Use of AI in the Math Classroom. *Mathematics Teacher: Learning and Teaching PK-12*, *117*(3), 236-241.

Ransom, M., & Manning, M. (2013). Teaching Strategies: Worksheets, Worksheets, Worksheets. *Childhood Education*, *89*(3), 188–190. https://doi.org/10.1080/00094056.2013.792707

Sawyer, A. G. (2024). Artificial Intelligence chatbot as a mathematics curriculum developer: Discovering preservice teachers' overconfidence in ChatGPT. *International Journal on Responsibility, 7(1),* 1–28.

Sawyer, A. G., Dredger, K., Myers, J., Barnes, S., Wilson, R., Sullivan, J., & Sawyer, D. (2020). Developing teachers as critical curators: Investigating elementary preservice teachers' inspirations for lesson planning. *Journal of Teacher Education*, *71*(5), 518–536. https://doi.org/10.1177/0022487119879894

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School*, *3*(4), 268–275.

Marcus, G. (2023). "Math is hard" – If you are an LLM – and why that matters. *Marcus on AI.* https://garymarcus.substack.com/p/math-is-hard-if-you-are-an-llm-and

Marr, B. (2023). 'A short history of ChatGPT: How we got to where we are today.' *Forbes, 19 May.* , page accessed 1.10.2024.

May, E. (2020). "Mathematical Proof | Definition, Parts & Example', Study. https://study.com/academy/lesson/mathematical-proof-definition-examples-quiz.html#:~:text=they%20are%20equivalent.-,Lesson%20Summary,definition%2C%20theorem%2C%20or%20postulate.

OpenAI (2023) 'Privacy policy.' *OpenAI*. 30, June, https://openai.com/policies/privacy-policy, page accessed 1.10.2024.

Wu, G. (2023). "8 Big Problems With OpenAI's ChatGPT." *MakeUseOf.* 17, September. https://www.makeuseof.com/openai-chatgpt-biggest-probelms/ page accessed 1.10.2024.