

# Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning

Sai Gattupalli<sup>1,†</sup>, Will Lee<sup>1,†</sup>, Danielle Alessio<sup>1,†</sup>, Danielle Crabtree<sup>1,†</sup>, Ivon Arroyo<sup>1,†</sup> and Beverly Woolf<sup>1,†</sup>

<sup>1</sup>University of Massachusetts Amherst, Amherst, MA 01003, USA

## Abstract

Despite the potential and emerging applications of large language models (LLMs) for education, little is known about their effectiveness in learning. Similarly, educators' preferences and perceptions on the utility of LLMs have received limited to no attention. Hence, we conducted an exploratory study to investigate pre-service teachers' perceptions (N=33) with respect to the possible utility of LLMs (such as GPT4) in online mathematics education. Our initial quantitative and qualitative findings indicate that while human-created mathematical contents - especially visuals - are still preferred, transformer-generated walk through, instructions, and guidance are helpful as tutoring in math problem-solving. Implications and future directions are also discussed.

## Keywords

Large Language Models, Evaluations, Intelligent Tutoring Systems, Mathematics

## 1. Introduction

In a recent "Dear Colleague" letter from the National Science Foundation [1], the agency stresses the importance of rapid research at the intersections of education and large language models (LLMs), both in informal and formal education settings. The letter underscores the hidden value, potential, and benefit of transformer-based LLMs [2] in education where they are designed to comprehend human language but might also be applied to support and tutor students.

In engaging K-12 students in math learning, effective pedagogical strategies and techniques often center around understanding students' learning goals and motivations [3]. These strategies could be further enhanced through the application of LLMs in maths education, as their potential to serve as "scaffolding" tools becomes increasingly apparent [3][4][5]. They can aid young learners in navigating math learning hurdles and streamlining their educational journey, directly responding to their learning needs and aspirations.

One example of this concept is MathSpring (MS), an intelligent online tutoring system developed by researchers from the University of Massachusetts Amherst and funded by the NSF. The platform aids students in their math problem-solving practice, reinforcing each problem


---

*AIED2023 Empowering Education with LLMs - the Next-Gen Interface and Content Generation, July 07, 2023, Tokyo, Japan*

✉ sgattupalli@umass.edu (S. Gattupalli); williamlee@cs.umass.edu (W. Lee); alessio@umass.edu (D. Alessio); dcrabtree@umass.edu (D. Crabtree); ivon@cs.umass.edu (I. Arroyo); bev@cs.umass.edu (B. Woolf)

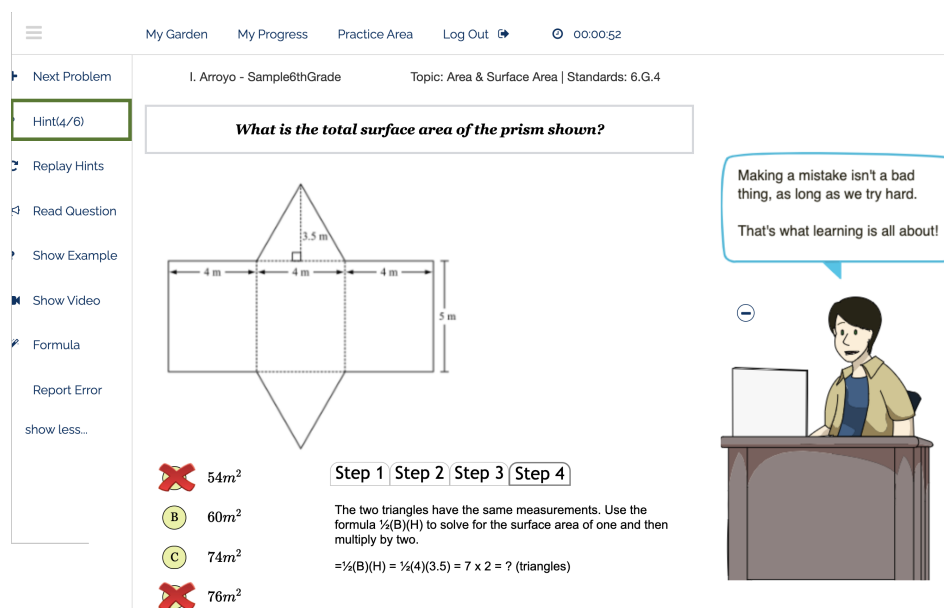


© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

with carefully crafted hints created by researchers with backgrounds in mathematics and education. However, the hint creation process currently requires significant time and expertise, and poses challenges related to human resources. The complexity of this process may contribute to burnout among teachers and researchers [6], underlining the need for methods to alleviate this burden.

In this light, integrating LLMs could offer a valuable solution, creating a responsive teaching strategy that supports both student engagement and teacher sustainability. LLMs could streamline the hint-creation process, in ways that help reduce the workload of educators while maintaining, or even enhancing, the quality of support provided to students. In turn, this could lead to a more engaging math classes which are capable of flexibly responding to K-12 students' learning goals and motivations.



**Figure 1:** The MathSpring online tutor's Practice Area interface, featuring the "Hints" button on the left, which offers textual hints accompanied by audio explanations. The platform also provides access to worked-out examples, tutorial videos, and formulas. Jake, the learning companion displayed on the right, encourages student perseverance and underscores the value of effort, especially when mistakes are made.

LLMs such as OpenAI's ChatGPT [7] emerge as potential game-changers, especially in K-12 math education. ChatGPT produces human-like explanation in text format and has shown efficacy across various sectors, including education, law, and medicine [8][9][10]. It also has potential in various learning tasks; including answering student math queries, crafting math hint narratives, summarizing math problems, and acting as a virtual math tutor [11]. An enhanced iteration of ChatGPT, based on GPT-4, was introduced in March 2023, is the focus of our work and showcases even more advanced natural language generation, understanding, and learning capabilities [7]. GPT-4's abilities are transforming not only learning but also higher education research [8]. Its knack for answering queries, providing explanations, and assisting in math

problem-solving makes it an asset in creating interactive learning experiences. Our work aims to explore the potential and effectiveness of GPT-4 produced hints for MS online tutor, in an attempt to make math learning more personalized and affective learning to aid learners (Grades 4 and up) in their math-problem-solving efforts.

In this exploratory study, the term "transformer-generated hints" are suggestions produced via prompt engineering techniques [12]. These hints can significantly benefit students grappling with mathematical word problems, which is a key skill often found challenging [13]. The term "human-crafted hints" refer to the hints created by real-humans. Building on [14]'s assertion that "automatically generating high-quality step-by-step solutions to math word problems has many applications in education," we stress the importance of teacher involvement in reviewing hints, irrespective of their origin – human-crafted or transformer-generated. We think a blend of transformer-generated hints and teacher input can promote learning while expediting the hint-creation process. This is essential for maintaining the reliability and effectiveness of MS, or any online learning platform.

## 2. Research Questions and Contribution of the Study

The value of this study lies in its dual focus: the advancement of online math education and the exploration of LLMs in educational contexts. As LLMs are gaining momentum for their potential to bolster learner engagement, academic achievement, and student success, little evaluation has been conducted. Hence, the outcomes of this study are primed to serve as a roadmap for educational stakeholders. This includes administrators and policymakers, assisting them to discern the potential advantages, constraints, and practical facets of integrating LLMs into online mathematics learning.

Moreover, by collecting the perceptions of pre-service teachers towards transformer-generated hints (Ht) compared to human-crafted hints (Hm), we aim to uncover preferences that could potentially inform future enhancements and integration of LLM-based tools within mathematics education. This inquiry thus proposes three contributions:

- Implementation of transformer-generated hints for K-12 mathematics education
- Assessment of teachers' perceptions and preferences regarding LLMs-generated hints, and their ensuing pedagogical implications
- The combined expertise in education, computer science, artificial intelligence, and math education, to investigate the utility of LLMs in math education.

Central to our study is our primary research question (RQ), which delves into the perceptions of pre-service teachers:

**RQ:** How do pre-service teachers perceive the effectiveness and appropriateness of transformer-generated hints in comparison to human-crafted hints?

Addressing this RQ allows us to underline the possible influences of transformer-generated hints on student learning.

### 3. Related Work

Our study serves as a initial effort for understanding the implications of LLMs in math education, particularly in relation to transformer-generated hints and their prospective role in elevating teaching and learning experiences. Radford [15] proposed the idea of applying transformers [2] by using semi-supervised approach where a natural language model using a large corpus of unlabelled text would be trained. In learning sciences, similar language models have been proposed and deployed for mathematical learning. For instance, MathBERT, introduced by [16], was trained using large volumes of mathematics-related text from K-12 and college-level mathematics textbooks, open-source course syllabuses, and research papers in mathematics. Griffith et al. [17] undertook quantitative experiments to compare different transformer-based neural networks for solving mathematical word problems, while other studies proposed using transformer models for solving differential equations [18], performing reasoning [19], and proofing [20].

However, these prior efforts have not tested their prototypes with real students and educators for feedback, leaving a crucial element unexplored. Moreover, many of these systems provide solutions without guiding students with hints, which is a significant aspect of learning.

Our work extends experiments conducted by [21] where the focus was on investigating the mathematical capabilities of ChatGPT, by treating ChatGPT as an assistant to professional mathematicians by posing various use cases such as question answering and theorem searching. Our approach is similar in ways that we aim to gather insights on the effectiveness and acceptance of LLMs-generated hints in our MS online tutoring system. The target demographic is undergraduate students preparing to become K-12 teachers and education professionals in the US.

### 4. Methodology

Participants (N=33) in this study are undergraduate students enrolled in education courses at a Northeastern university. All participants aspire to become K-12 teachers and education professionals. The participants are active members of the Education Club, a student-led group committed to nurturing connections among educators and the wider community. Their perceptions and responses towards both human-crafted and transformer-generated hints were gathered within an hour-long session during the club's weekly meetings. Given this is an exploratory study, we did not collect any demographic data.

We adopted a mixed-methods approach in this exploratory study, striking a balance between the quantitative analysis of hint counts and the qualitative insights derived from open-ended responses. Our qualitative analysis employs the grounded theory approach, which guides the thematic dissection of the participant responses.

Five multiple-choice math word problems based on Common Core standards were randomly selected from the MS SQL production database and presented to participants via a Qualtrics survey. Participants gave their preference between two distinct hint variants (one human-created by education professionals and the other generated by GPT-4) and answered a "Why did you choose this hint variant?" open-response question. All survey questions have been

made available at <https://osf.io/t84v7/>.

For definitions, human-crafted hints ( $H_m$ ) embody the collective wisdom of math teachers, research scientists, and university professors. These hints emphasize pedagogical strategies that enhance individual learning and problem-solving abilities. However, as discussed above, their creation is a time-consuming process, requiring the skillful expertise of educators. In contrast, the transformer-generated hints ( $H_t$ ) were produced using the GPT-4, utilizing prompts tailored from a Prompt Engineering GitHub repo [22]. This approach aimed to simulate the instructional strategies of an experienced US math teacher.

The prompts fed to GPT-4 were designed to place the model in the role of a math teacher with a decade's worth of experience teaching students from a variety of economic, ethnic, cultural, and language backgrounds.

Figure 2 shows a sample MS question that all participants responded to, along with the hint variants.

#### 4.1. Data Collection

Every participant responded to the Qualtrics survey by indicating their preferred hint variant and their rationale behind choosing it. Notably, the participants were blind to whether the hints were human-created or transformer-generated, which is an approach taken to eliminate any potential biases and preconceived notions about advanced technologies or GPT-based services in general. The collected data was organized into a spreadsheet, which provided the basis for our ensuing analysis detailed in the below section. The full dataset has been made available at <https://osf.io/t84v7/>.

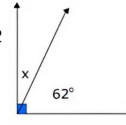
### 5. Results

#### 5.1. Analysis

The analysis of the participants' hint preferences was carried out via a histogram representation for clarity and ease of understanding. In addition, we employed a pre-trained BART language model from the Hugging Face API [23] to compile and interpret participants' written responses, the insights from which are further discussed in the Discussion and Conclusion section.

To comprehensively categorize the collected written responses, we used a pre-trained BERT model in conjunction with k-means clustering. This allowed us to discern prominent themes

Q: Find the value of  $x$  in a right angle with a line at  $62^\circ$  degrees, where  $x$  is the complementary angle.



**Hint 1:** There are 2 variants of a hint below. Choose the one you feel is the most effective as the first hint.

Complementary angles are two angles with a sum of  $90^\circ$ . Two adjacent complementary angles form a right angle.

Complementary angles are two angles that add up to  $90^\circ$ . When they are adjacent to each other, they form a right angle. Think of a corner or an L-shape as an example.

**Hint 2:** The student still did not understand how to solve the problem, and asks for a second hint. Which of the following is more effective? (or most beneficial, effective, etc)

From the diagram, we see that these are complementary angles.

Look at the diagram closely. Can you identify the right angle and the two angles that make it up? If so, then you know these are complementary angles.

**Hint 3:** The student was unsure about the steps and processes to come up with a solution. The student decided to ask for a third hint. Which of the following is more effective? (or most beneficial, effective, etc)

Therefore,  $x+62^\circ=90^\circ$

Now that you've identified the complementary angles, let's set up an equation to find the value of  $x$ :  $x + 62^\circ = 90^\circ$ . Remember, the sum of complementary angles is always  $90^\circ$ .

Please explain why you chose this variant.

A large, empty rectangular text box with a light gray background and a thin border, intended for the participant to provide an explanation for their chosen hint variant.

Figure 2: Participant survey question. Participants chose their preferred hint and explained why they preferred it.

associated with the three types of hints under consideration: human-crafted hints ( $H_m$ ), transformer-generated hints ( $H_t$ ), and mixed-result ( $H_{mt}$ ). A detailed summary of these preferences can be found in Table 1. We conducted our analysis using BART and BERT models, along with k-means clustering and visualization, all within Google Colab. Our notebooks have been made available online.<sup>1</sup>

## 5.2. Findings

The results, as illustrated in Table 1, are segmented into three sections. We first present the instances where participants showed a preference for human-crafted hints (Q1 and Q4). This is followed by an examination of the instances where participants favored transformer-generated hints (Q2 and Q3). Lastly, we delve into the case where the results were mixed (Q5).

Each question’s results were organized according to the hint variant participants preferred: human-crafted or transformer-generated. Therefore, Q1 and Q4, which both indicated a preference for human-crafted hints, are discussed together. Similarly, we grouped Q2 and Q3, as they shared a favorability for transformer-generated hints. The analysis of Q5, which revealed mixed preferences among participants, is in the discussion section.

**Table 1**  
Preferred Hint Choice from Participant Responses

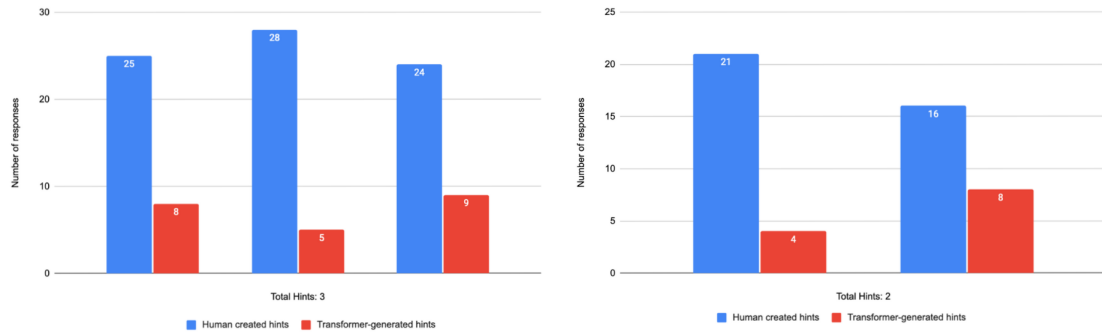
Question ID	Preferred Hint Choice
Q1	$H_m$ (Human-crafted)
Q2	$H_t$ (Transformer-generated)
Q3	$H_t$ (Transformer-generated)
Q4	$H_m$ (Human-crafted)
Q5	$H_{mt}$ (Mixed-result)

### 5.2.1. Human-Crafted Hints ( $H_m$ )

Participants preferred human-crafted hints over transformer-generated hints for Q1 and Q4, see Figure 3. The *Visual Cue* themes from the qualitative data (responses) that emerged from Grounded theory and thematic analysis are *Explanation with Example*, *Visualization* and *Explanation in Detail* and are displayed in Table 2 along with sample participant utterances.

<sup>1</sup>[https://github.com/wlee-umass/AIEDLLM\\_Workshop](https://github.com/wlee-umass/AIEDLLM_Workshop)

### Preferred Human-crafted hint choice Q1 and Q4



**Figure 3:** Participant responses for Q1 and Q4 were grouped into themes for human-crafted. Here, we see that the human-created hints are preferred over the transformer hints.

**Table 2**

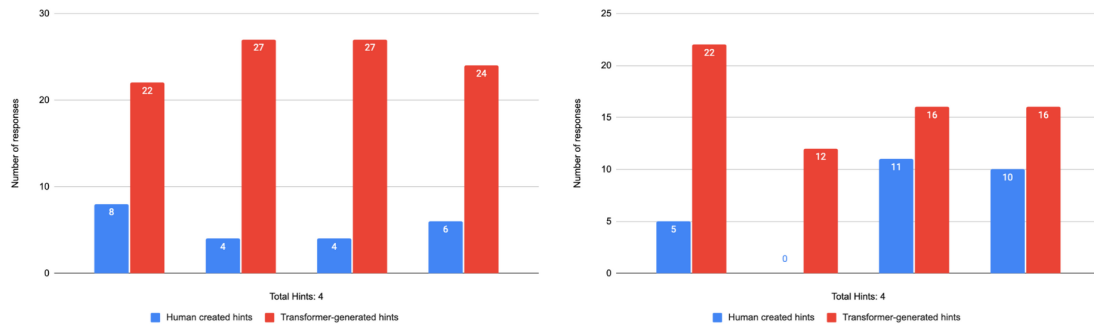
Using a pre-trained BERT [23], written responses of Q1 and Q4 were converted into embeddings. K-means clustering was used to group these embeddings into themes. A selection of top themes was extracted from the clusters of themes.

Topic Cluster	Theme from Human-Created Hint - Visual Cues
1: Explanation with Example	<ul style="list-style-type: none"> <li>• "I like being able to see the explanation in pictures/numbers more than words. I get a bit lost in the words version."</li> <li>• "I like how in this one the expression is written out and how to find the multiples of each number."</li> </ul>
2: Visualization	<ul style="list-style-type: none"> <li>• "The visual! Much more helpful than the words."</li> <li>• "[T]he picture helps."</li> </ul>
3: Explanation in Detail	<ul style="list-style-type: none"> <li>• "Explains why it works and substitute the original equation."</li> <li>• "If a student needs an additional hint, they may need a little more information, which is why I chose the first option."</li> </ul>

#### 5.2.2. Transformer-Generated Hints ( $H_t$ )

Participants preferred transformer-generated hints over human-crafted hints for Q2 and Q3, see Figure 4. The themes from the qualitative data (responses) that emerged from Grounded theory and thematic analysis are *Explanation Through Connection*, *Guidance* and *Simple Walkthrough* and are displayed in Table 3 along with sample participant utterances.

### Preferred transformer-generated hint choice Q2 and Q3



**Figure 4:** Participant responses for Q2 and Q3 were grouped into themes for transformer-generated created. Here, we see that the transformer-generated hints are preferred over the human-created hints.

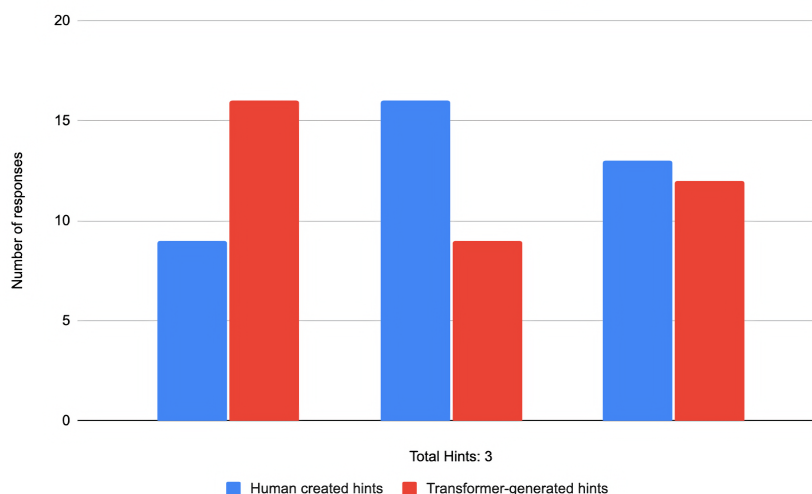
**Table 3**

Using a pre-trained BERT, written responses of Q2 and Q3 were converted into embeddings. K-means clustering was used to group these embeddings into themes. A selection of top themes was extracted from the clusters of themes

Topic Cluster	Theme from Transformer-Generated Hint
1: Explanation Through Connection	<ul style="list-style-type: none"> <li>• "The language is more descriptive ie the explanation that a right angle forms an L shape."</li> <li>• "I feel like when it comes to applying concepts in an equation format it becomes more confusing to not elaborate on how equations are formed or why they appear the way they do. Even if the previous two hints lead up to the equation itself, having this hint appear like this might draw the two hints into a full circle with this elaboration, because if the child struggles with just the first two hints, they might need more than just given steps, and rather, scaffolding alongside with the information provided."</li> </ul>
2: Guidance	<ul style="list-style-type: none"> <li>• "This coincides with the first question and is closer to what they have most likely been taught!"</li> <li>• "[T]his guides students through the problem."</li> </ul>
3: Simple Walkthrough	<ul style="list-style-type: none"> <li>• "[L]eads the student into the correct 1st step."</li> <li>• "[M]akes it very easy for a student to understand what they should be doing."</li> </ul>



## Mixed-hint choice Q5



**Figure 5:** Participant responses for Q5 distributed according to human created or transformer created. Here, we see that the participants had a mixed-preference for both human-crafted and transformer-generated hints.

### 5.2.3. Mixed - Human-Crafted Hints/Transformer-Generated Hints ( $H_{mt}$ )

Participants had a mixed preference for both the human-created and transformer-generated hints for Q5, see Figure 5. Themes that arose from the participants' responses to why they preferred certain hints were ease of understanding and comprehension. Example of utterances from these themes include "This is less dense to read and easier to understand" and "This explains the process of converting a fraction to a decimal which is important for the students comprehension instead of just giving them the answer."

## 6. Discussion and Conclusion

When it comes to human-crafted hints, one trend we observed from the thematic analysis is that participants preferred the teacher-crafted visual cues ( $H_m$ ). This is evident from the summaries of participants' written responses. For example, we identified that participants in this category preferred "seeing the numbers (is) easier than just seeing the words... and this is more helpful because it has both a verbal component and a visual component." The emphasis on visual cues was perceived as beneficial because they complemented the textual information. This observation substantiates the pedagogical importance of visual aids in enhancing comprehension and reinforcing concepts.

In the case of transformer-generated hints ( $H_t$ ), our analysis revealed that participants favored detailed, step-by-step instructions, and clear narratives. Participants appreciated the guidance towards the subsequent stages of problem-solving. For example, from the summarization we conducted, participants indicated that "I liked being talked through the problem. I like the

inclusion of “think about the letter L’. This has a more readable, in depth explanation. This helps explain what a complementary angle is AND gives a sufficient example.” This feedback suggests that future applications of LLMs in intelligent tutoring systems should simulate a conversational problem-solving process, while allowing for interactive queries from the students. An interesting observation from a consultant K-12 teacher suggested that GPT-4 generated hints might be too lengthy and detailed for young learners with shorter attention spans. The transformer-generated hints were deemed beneficial by the participants despite the lack of visuals, attributed mainly to their explanatory verbiage.

One interesting finding is a mixed result of Q5. As seen in Figure 5, participants rated human-crafted hints and transformer-generated hints almost equally. One reason we observed is that the human-crafted and transformer-generated hints were framed similarly, resulting in a balanced preference and hence the mixed result. Unlike the other four problems, we speculate that the use of a third-person narrative in Q5 may have influenced the participants to perceive the hints as less personally relevant in their math problem solving. Therefore, the resemblance between the two hint variations might have led the participants to not pay as much attention to the minute differences between each hint variants.

There are two limitations in this exploratory study. One limitation is that due to time constraint, we only selected five mathematical problems - although the problems spanned a wide range of topics. Second limitation is the only choice of OpenAI’s GPT-4 model for generating hints. We believe exploring and comparing other pre-trained decoder based models - or even fine-tuning a model - in the future is essential to generate readable and interpretable hints.

Looking ahead, our plan is to evaluate hints generated by LLMs using the Flesch Reading Ease Formula [24]. This tool will help us understand the effectiveness of generated hints to further explain participants’ responses. This is important because students with reading disabilities may find the substantial reading involved with LLMs challenging. Although our current work does not focus on translations, we believe LLMs can facilitate the scaling of auto-translation for generated hints on-demand. The ability to translate hints, math contents, and be able to provide emotional support in different languages on-demand is crucial if LLMs are to accommodate bilingualism [25], and non-native English Language Learners (ELL).

We foresee using LLMs in crafting digital learning companions [26], with varied skin tones and languages, to be embedded into learning platforms such as MS, in our attempt to foster a more personalized learning experience for all learners. We believe our collected log data from past experiments, such as math mastery, time spent, and correctness, can be utilized to fine-tune or introduce new layers into a scalable LLM model. The outcome of this adaptation may serve as an advanced intelligent tutor capable of guiding and supporting students interactively, irrespective of their background.

To conclude, our study utilizing OpenAI’s GPT-4 LLM aimed to explore the perspectives and preferences of pre-service teachers concerning the efficacy of transformer-generated hints. This research provides broad implications, one being the potential development of more personalized and intelligent features and tools that support all MS users. We aim to further the understanding of personalized learning experiences and the role of LLMs in shaping education.

## References

- [1] Dear Colleague Letter: Rapidly Accelerating Research on Artificial Intelligence in K-12 Education in Formal and Informal Settings (nsf23097) | NSF - National Science Foundation — nsf.gov, <https://www.nsf.gov/pubs/2023/nsf23097/nsf23097.jsp?org=NSF>, 2023. [Accessed 28-May-2023].
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [3] I. Arroyo, B. Woolf, W. Burelson, K. Muldner, D. Rai, M. Tai, A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect, *International Journal of Artificial Intelligence in Education* 24 (2014). doi:10.1007/s40593-014-0023-y.
- [4] J. Anghileri, Scaffolding practices that enhance mathematics learning, *Journal of Mathematics Teacher Education* 9 (2006) 33–52.
- [5] A. Bakker, J. Smit, R. Wegerif, Scaffolding and dialogic teaching in mathematics education: Introduction and review, *ZDM* 47 (2015) 1047–1065.
- [6] S. Woods, J. Sebastian, K. C. Herman, F. L. Huang, W. M. Reinke, A. M. Thompson, The relationship between teacher stress and job satisfaction as moderated by coping, *Psychology in the Schools* (2023).
- [7] OpenAI, Gpt-4 technical report, 2023. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [8] H. Gimpel, K. Hall, S. Decker, T. Eymann, L. Lämmermann, A. Mädche, M. Röglinger, C. Ruiner, M. Schoch, M. Schoop, et al., Unlocking the power of generative ai models and systems such as gpt-4 and chatgpt for higher education (2023).
- [9] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).
- [10] D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam, Available at SSRN 4389233 (2023).
- [11] Y. Wardat, M. A. Tashtoush, R. AlAli, A. M. Jarrah, Chatgpt: A revolutionary tool for teaching and learning mathematics, *Eurasia Journal of Mathematics, Science and Technology Education* 19 (2023) em2286.
- [12] GitHub - f/awesome-chatgpt-prompts: This repo includes ChatGPT prompt curation to use ChatGPT better. — [github.com](https://github.com/f/awesome-chatgpt-prompts), <https://github.com/f/awesome-chatgpt-prompts>, 2023. [Accessed 29-May-2023].
- [13] R. Misquitta, A review of the literature: Fraction instruction for struggling learners in mathematics, *Learning Disabilities Research & Practice* 26 (2011) 109–119.
- [14] J. He-Yueya, G. Poesia, R. E. Wang, N. D. Goodman, Solving math word problems by combining language models with symbolic solvers, *arXiv preprint arXiv:2304.09102* (2023).
- [15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).
- [16] J. T. Shen, M. Yamashita, E. Prihar, N. Heffernan, X. Wu, B. Graff, D. Lee, Mathbert: A pre-trained language model for general nlp tasks in mathematics education, *arXiv preprint arXiv:2106.07340* (2021).
- [17] K. Griffith, J. Kalita, Solving arithmetic word problems with transformers and preprocessing of problem text, *arXiv preprint arXiv:2106.00893* (2021).

- [18] G. Lample, F. Charton, Deep learning for symbolic mathematics, arXiv preprint arXiv:1912.01412 (2019).
- [19] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, et al., Solving quantitative reasoning problems with language models, arXiv preprint arXiv:2206.14858 (2022).
- [20] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al., Training verifiers to solve math word problems, arXiv preprint arXiv:2110.14168 (2021).
- [21] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, J. Berner, Mathematical capabilities of chatgpt, arXiv preprint arXiv:2301.13867 (2023).
- [22] F. K. Akın, Awesome chatgpt prompts, GitHub, 2023. URL: <https://github.com/f/awesome-chatgpt-prompts>, online; Accessed: 2023-05-23.
- [23] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
- [24] J. N. Farr, J. J. Jenkins, D. G. Paterson, Simplification of flesch reading ease formula., *Journal of applied psychology* 35 (1951) 333.
- [25] D. Alessio, B. Woolf, N. Wixon, F. R. Sullivan, M. Tai, I. Arroyo, Ella me ayudó (she helped me): Supporting hispanic and english language learners in a math its, in: *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II* 19, Springer, 2018, pp. 26–30.
- [26] B. P. Woolf, I. Arroyo, K. Muldner, W. Burleson, D. G. Cooper, R. P. Dolan, R. Christopherson, The effect of motivational learning companions on low achieving students and students with disabilities., in: *Intelligent Tutoring Systems (1)*, 2010, pp. 327–337.