Recent Advances of Artificial Intelligence in Math Education

Mitchell Piehl

March 14, 2025

Abstract

Recent advances in artificial intelligence have affected almost every area of our daily lives. These advances help various processes to be more productive, efficient, and cost-reducing. However, are these advances always beneficial? Is the increasing availability of artificial intelligence models that can more accurately find answers and solve problems than ever before helpful for today's students? This paper will describe the recent advances in artificial intelligence and analyze how it affects students and teachers in educational settings, specifically in mathematics.

1 Introduction

The rapid growth of Artificial Intelligence applications in recent years has had a massive effect on the daily lives of individuals, including students and teachers. These applications give rise to more creative, innovative thinking and more direct and applicable help, which helps students get much more individualized assistance. However, these capabilities could be just as harmful as they are beneficial. This paper aims to analyze the balance between the benefits and harms of AI for teachers, professors, and developers while considering how to address this rapidly growing field.

I chose this topic for this paper because it is valuable for all researchers and developers to analyze the impact of their work to ensure it supports the common good and is ethically beneficial for society. As I continue my education into graduate school to obtain my PhD in a related field to this paper, it is essential to develop an appropriate understanding of the ethical impacts of my work and how I can enhance the common good.

1.1 What is AI?

Before we can analyze the effects of artificial intelligence, it is essential to understand what it is. While AI can be difficult to define, it is generally agreed that AI is a process of producing or mimicking human intelligence in machines or computers. There are various fields and applications of AI, including knowledge representation, heuristic search, planning, expert systems, machine vision, machine learning, natural language processing, software agents, intelligent tutoring systems, and robotics [FR14]. While all of these fields are relevant in math education today, this paper will only address the effects of the fields of natural language processing and robotics.

Natural Language Processing (NLP) is a subfield of AI that includes understanding and generating natural language. Today, NLP models can be found everywhere, but the most popular form of NLP is through transformer-based large language models such as ChatGPT. Recent developments in this subfield affecting math education include tutor assistants, math problem solvers, and more.

Robotics has historically been a subfield of mechanical engineering. However, today, the use of Artificial Intelligence within robotics has created a significant overlap between the two fields. It has allowed students to use robots in educational settings to apply mathematical and engineering knowledge. This paper specifically talks about these two fields because these are the most popular ways AI is used in math education settings. Robots used to be the most common use of AI in mathematics classroom settings; however, natural language processing tools have since taken the lead and look to continue pushing ahead.

2 Historical Developments of Artificial Intelligence

Artificial Intelligence is not a recent invention by any means. The term Artificial Intelligence comes from the 1956 Dartmouth conference, where John McCarthy and his team of Marvin Minsky, Herbert Simon, Allen Newell, and others came together to share ideas and solve AI. However, McCarthy and his colleagues quickly discovered the difficulties of AI development. Not much was gained from the Dartmouth conference besides creating the name Artificial Intelligence, which is still a name that some experts in the field do not enjoy. However, now Artificial Intelligence was an official field in computer science, marking the start of the golden age of AI.

2.1 The Golden Age

The golden age of AI saw many attempts at problem solvers that were quite good in comparison to the computing power of that age. One of the first significant developments during the golden age was the General Problem Solver, or GPS, created by Herbert Simon, John Clark Shaw, and Allen Newell in 1959. This general problem solver was created to mimic human problem-solving. The program demonstrated abilities of memory, organization, and planning [NSS59]. Another significant development that came from the golden age of AI was Eliza. Eliza was a natural language processing program created by Joseph Weizenbaum that attempted to imitate conversations with humans by identifying keywords in the input text and providing a pre-created output text [Wei66]. A third significant development of Artificial Intelligence during the golden age was the SHRDLU system in 1971. The SHRDLU was one of the most acclaimed systems of the golden age, as it demonstrated novice problem-solving and natural language understanding. Other significant developments of the golden age include an expert system called MYCIN, created in Stanford's Lab in the 1970s, intending to be a doctor's assistant and actually could outperform humans in essential problems; the R1/XCON system was another AI system that saved it's company around 40 million dollars in the early 70's, and the DENDRAL project was the first large-scale program to explore automation of the acquisition of task-specific knowledge in the field of organic chemistry.

Despite these great successes in a newly developing field, these systems had many limitations. They often had very specific domain knowledge and zero capabilities outside of their field. Although these systems were impressive, they fell short of expectations of the time, which led to an "AI winter," which limited funding in the field and decreased the amount of growth seen.

2.2 AI beats humans

It wasn't long after the birth of Artificial Intelligence that it started surpassing humans in various activities. The first area where AI became superior to human performance was checkers. In 1959, Arthur Samuel created the first machine-learning algorithm through his checker-playing program. It took just a couple of months after making his program for it to be so good that Samuel could not beat it anymore [Sam59]. The success of AI in checkers was impactful but didn't take many by surprise; after all, checkers is a reasonably easy game with a limited number of move possibilities, making it manageable to master. The first big success of AI over humans came in 1997, when IBM's Deep Blue beat the world champion, Gary Kasparov, in a six-game match. Chess was a more complicated game to master, and at the time, it was understood to require large amounts of intelligence. Deep Blue beat Kasparov using a traditional AI game-playing algorithm, a game tree with various search algorithms, and a complex evaluation function [CHJH02]. This win was the first time people considered machines to have human-like intelligence. That same year, AI again showed its superiority over humans, this time in mathematics. Before 1997, the Robbins conjecture was a famously complex math problem, unsolved by mathematicians for over 50 years. Still, an Artificial Intelligence automatic theorem prover called Equational Prover, or EQP, successfully proved this theorem, already showing some significant advantages of artificial intelligence in mathematics [McC97]. In 2011, IBM saw another big success with their artificial intelligence systems when Watson, their AI system built to answer jeopardy questions, outperformed the best in the world on the nationally televised show Jeopardy. Watson used a natural language processing algorithm to query 200 million content pages while replying appropriately. One of the biggest wins of artificial intelligence came in 2015 when Google's AlphaGo beat the best Go player in the world using artificial neural networks. This win was so impactful because Go is a very complex game, with over 250 possible moves from an average board position, and it is nearly

impossible to create an evaluation function because of its complex nature. The best humans have to rely on intuition to play the game. Using deep Q learning, AlphaGo not only managed to win against the world champion, but it played so well it taught players new moves that were so advanced competitors thought they were mistakes at the time of play [M+19]. The DeepMind team at Google continued to do impressive work as they developed a more general player called AlphaZero. AlphaZero learned to play a number of different Atari games and performed at super-human levels in many of them. AlphaZero's capability of winning a number of games was the first time neural networks were able to be used in multiple games, demonstrating increased generality and the large potential for further growth.

Again, we see many successes in the field of AI, as AI is developing at a rapid pace. However, these successes are still limited. These AI systems are called Narrow AI systems, meaning they are only good at one specialized task and fail to be successful outside of their trained domain. For example, even though Deep Blue could beat the greatest in the world in chess, it could not even finish a game of checkers. Additionally, even small changes to the trained domains of the agents substantially lower their performance. For example, DeepMinds AlphaZero can play pong at a super-human level, but as soon as the paddle is raised by just one pixel, the performance goes down to worse than that of the average human. This lack of generality made it nearly impossible for these programs to be helpful in constantly changing classroom-based settings.

2.3 The Rise of Transformer based LLMs

Google's introduction of the transformer in 2017 was possibly the biggest step in making AI more generalized in its history. The transformer is a type of deep neural network that uses forms of multi-headed attention to process input sequences in parallel, thus becoming significantly more efficient than the previous Recurrent Neural Network-based models [VSP+17]. This new transformer architecture could then be used as the base of new large language models that are significantly more accurate and effective than before.

Large-language models are algorithms that take a sequence of input words (today translated into word vectors, thanks to Google's Word2Vec) and predict an output based on training data. LLMs have existed since 1948 when Shannon used the Markov chain framework to create an n-gram large language model, but today's transformer-based LLMs are significantly more advanced. In addition, the availability of the worldwide web content provides more extensive training data for these large language models than ever before.

The first global success of a transformer-based LLM came from OpenAI's ChatGPT, released in 2022. This model could converse with humans just as another human would. It could answer questions on a wide range of topics and speak in better English than most native English speakers. Since the release of ChatGPT, these models have found their way into the lives of almost everyone in the digital world. Businesses use these AI models to help chat with customers; computer scientists can use them to help them with their projects, they can create poetry, new songs, and music, and most importantly for this paper, these AI models can complete students' homework, help students study for exams, and even grade papers.

2.4 LLMs solving math

Although these models are extremely powerful, many aspects of them still cause them to struggle with math problems. For example, these models use probabilities to determine outputs to create unique responses, but most math problems do not involve any sort of probability or change in answers. This causes decreases in accuracy on top of their already low accuracy. As models continue to get larger and better, they naturally become better at solving math equations. For example, GPT-4 solved math word problems from a common algebra dataset with nearly 35 percent higher accuracy than GPT-3.5. [AAA+23]. However, other ways exist to improve performance without larger and more advanced models.

Researchers have also developed more advanced ways to use these AI models to solve math problems through prompt engineering to increase accuracy. A simple yet effective version of this is called Chain of Thought (CoT) prompting, which asks the large language model to explain its thoughts while solving a math problem [WWS+22]. This process is similar to how students can more accurately solve math problems when they reason through them. It is also beneficial for students because the

provided explanation helps students understand how the math problem is being solved. Another popular technique used to increase accuracy is few-shot prompting. This technique gives the model a few examples of similar problems being solved before asking the model to solve the problem. This idea is similar to how students are provided examples of how problems are solved before being asked to solve a problem themselves [BMR+20]. There are numerous other techniques used to increase accuracy such as self-consistency, asking the same question repeatedly then finding the mode of those answers [WWS+23], decomposition, asking the LLM to simplify the problem into parts before re-asking the model to solve [KTF+23], progressive-rectification prompting introduced in a paper called "Get an A in Math..." due to 90 percent accuracies in elementary level questions, which checks answers by re-asking the LLM the same question with a different value removed from the equation [WJS24], or even my own method developed that I developed recently that prompts the LLM to estimate the correct answer and then compares that answer to an answer generated by a symbolic solver. This new method, which I call Estimation Verification of Symbolic Solving, or EVoSS, produced accuracies averaging almost 90 percent on math word problem datasets that contain questions around the 8th-grade level [PWKK25].

As these solvers become more advanced at rapid speeds, they will soon be able to solve math problems at higher educational levels, becoming applicable and usable at all levels. This introduces many critical ethical questions: Do these solvers support or add to the common good? Do they help students more or hurt students more? How can educators use these tools to their advantage by providing students with easier-to-access help? These questions must be answered before these models become too integrated into students' lives.

2.5 AI Robots in Math

Another very popular and potentially less controversial field of Artificial Intelligence in mathematics education is robotics [bMHbS+22]. AI robots are the most popular way AI is used in classroom settings outside natural language processing AI tools. Robotics in math provides a clear and fun way for students to apply their knowledge in the real world; also, because they provide interactive feedback, using robotics helps students advance their reasoning abilities. Teachers have been able to use the programming of robots within classrooms to teach or apply mathematical concepts [CF21]. Robots are an excellent way of answering a common question in today's math classes; "When am I ever going to need to know this outside of class?" The advances in artificial intelligence have led to robots being more versatile and being used in a new variety of ways because robot programming can be more specific to the topic in class while AI handles the rest of the programming. Additionally, advancements in AI lead to more complex robots that will further increase students' attentiveness and positive emotions.

3 Impacts on Math Education

As mentioned many times in this paper, the use of AI has many potential impacts on math education. In this section, this paper will divide the good and bad effects before combining them and providing a final analysis.

3.1 The Pros

There are many pros to using artificial intelligence in classroom settings. For example, Generative AI tools can excel in providing students with on-demand problem-solving that is specific to their problem. These tools have also been integrated into tutoring systems that help students solve math problems without needing a teacher. [AYD25]. These tools do a great job of addressing the stress applied to teachers in large classrooms of students who cannot get to all the students promptly. One study concludes that Artificial Intelligence makes teaching and learning in math education more effective because it is exciting and creative and has made it easier for students to understand subjects [bMHbS+22]. Another study found that ChatGPT shifted 9th-12th grade students' attitudes, bolstered interests in mathematics, and improved perceived self-efficacy. Additionally, scores were clearly improved across assessments, quizzes, and problem-solving tasks [Pat23]. Overall, it has been found that the use of AI in classroom settings has a strong connection with an increase in positive academic emotions towards mathematics [EA23].

There are also numerous unique and beneficial ways students use AI to assist in their learning. For example, students can use these tools to create study guides for their upcoming exams, grade their homework assignments before final submission, which provides instant interactive feedback, or plan when they should study for each class based on their exam schedules. These are just a few examples that show the broadness of the advantages of these AI tools.

The field of robotics has also found many beneficial ways of improving students' education. For example, one study used robots to test and facilitate students' multiplication skills, which improved most students' overall math achievements [HHKvB21]. Another study found that using robots in classrooms provided a social environment that allowed students to build their mathematical knowledge and develop their proportional reasoning skills [CF18]. Overall, the rise of Artificial Intelligence has allowed teachers to use robots in creative and unique ways that help students learn in new settings and apply their knowledge.

3.2 The Cons

Despite AI's numerous positive impacts, negative impacts come along with these new tools as well. One disadvantage that all teachers know is that generative AI tools such as ChatGPT threaten the integrity of assignments and online exams. As mentioned numerous times, these AI tools can answer questions reasonably well, and students can use that to forge their work and create questions of authenticity in assignments and tests. This is bad for both parties. No teacher wants to read homework assignments answered by AI, and the students don't learn anything when AI does the homework for them. Another disadvantage of using AI is the creation of blind reliance on AI-generated answers. The ease of obtaining answers limits critical thinking skills. This blind reliance can increase the spread of misinformation and negatively impact research and education. Another significant negative impact that is considered less often is the potential biases and ethical implications embedded in the training data. If the data that is used to train these generative AI models contains underlying biases, it can generate discriminatory or unfair results [RW23]. Additionally, since these generative AI tools are 'black boxes, ' it is nearly impossible to see what they have learned until they are tested sufficiently. However, knowing what sufficiently means in this scenario is also almost impossible because tiny input changes can sometimes drastically impact the outputs of these models. This phenomenon is best demonstrated by adversarial attack examples, where the input to AI models is changed so slightly that they are often unperceivable to the human eye, yet the output is changed so drastically that it becomes completely unreliable. These numerous negative impacts of AI could affect students' mathematical understanding and classroom performance and their abilities to gather information, think critically, and act responsibly, which will cause classroom settings to change drastically.

4 Conclusion

While there are positives and negatives to using artificial intelligence in mathematics education, there appears to be a large amount of support for using AI. This is best demonstrated by a meta-analysis completed in 2022 that reviewed 21 articles and 30 independent samples on the effects of AI on elementary students' mathematics achievement and found 27 of the 30 found an overall positive impact, concluding the use of AI significantly improves elementary students' mathematical achievements [Hwa22]. However, this does not mean that the negative impacts should be forgotten about or ignored. To fully benefit from using AI tools and introducing AI in classrooms while mitigating the negative consequences, everyone will need to use safe use practices. Developers will need to ensure models are trained without biases, act predictably, are not not susceptible to adversarial attacks. Teachers will need to ensure exams cannot be forged by generative AI and find ways to appreciate the use of AI in assisting students with homework. Then, students must uphold their critical thinking abilities, not unquestioningly trust generative AI, and maintain integrity in their answers. While these criteria may seem far-fetched, if they are held, the use of AI in mathematics education can advance the common good by helping students and teachers in their educational settings.

References

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2023.

[AYD25] Liz A Awang, Farrah D Yusop, and Mahmoud Danaee. Current practices and future direction of artificial intelligence in mathematics education: A systematic review. *International Electronic Journal of Mathematics Education*, 20(2):em0823, 2025.

[bMHbS+22] Mohamed Zulhilmi bin Mohamed, Riyan Hidayat, Nurain Nabilah binti Suhaizi, Muhamad Khairul Hakim bin Mahmud, Siti Nurshafikah binti Baharuddin, et al. Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education*, 17(3):em0694, 2022.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[CF18] Shelli L Casler-Failing. Robotics and math: Using action research to study growth problems. *The Canadian Journal of Action Research*, 19(2):4–25, 2018.

[CF21] Shelli Casler-Failing. Learning to teach mathematics with robots: Developing the 't'in technological pedagogical content knowledge. *Research in Learning Technology*, 29, 2021.

[CHJH02] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[EA23] Mohamed Elsayed and Sahar Abdo. Applications of artificial intelligence and their relationship to spatial thinking and academic emotions towards mathematics: Perspectives from educational supervisors. *Eurasian Journal of Educational Research (EJER)*, (107), 2023.

[FR14] Keith Frankish and William M Ramsey. *The Cambridge handbook of artificial intelligence*. Cambridge University Press, 2014.

[HHKvB21] Johan F Hoorn, Ivy S Huang, Elly A Konijn, and Lars van Buuren. Robot tutoring of multiplication: Over one-third learning gain for most, learning loss for some. *Robotics*, 10(1):16, 2021.

[Hwa22] Sunghwan Hwang. Examining the effects of artificial intelligence on elementary students' mathematics achievement: A meta-analysis. *Sustainability*, 14(20):13185, 2022.

[KTF+23] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. Camera-ready version.

[M+19] Melanie Mitchell et al. Artificial intelligence: A guide for thinking humans. 2019.

[McC97] William McCune. Well-behaved search and the robbins problem. In *Rewriting Techniques and Applications: 8th International Conference, RTA-97 Sitges, Spain, June 2–5, 1997 Proceedings 8*, pages 1–7. Springer, 1997.

[NSS59] Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, page 64. Pittsburgh, PA, 1959.

[Pat23] Judelyn L Patero. Revolutionizing math education: Harnessing chatgpt for student success. *International Journal of Advanced Research in Science, Communication and Technology*, 2023.

[PWKK25] Mitchell Piehl, Dillon Wilson, Ananya Kalita, and Jugal Kalita. Solving math word problems using estimation verification and equation generation. *Available at SSRN 5166090*, 2025.

[RW23] Md Mostafizer Rahman and Yutaka Watanobe. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied sciences*, 13(9):5783, 2023.

[Sam59] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wei66] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[WJS24] Zhenyu Wu, Meng Jiang, and Chao Shen. Get an a in math: Progressive rectification prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19288–19296, 2024.

[WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[WWS+23] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. Camera-ready version.