Pedagogical Research
2024, 9(2), em0188
e-ISSN: 2468-4929
https://www.pedagogicalresearch.com

MODESTUM

Research Article

OPEN ACCESS

An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam

Nezihe Korkmaz Guler [1]*, Zeynep Gul Dertli [2], Elif Boran [1], Bahadir Yildiz [2]

[1] Çukurova University, Adana, TÜRKİYE
[2] Hacettepe University, Ankara, TÜRKİYE
***Corresponding Author:** nezihekorkmazguler@gmail.com

Citation: Korkmaz Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research, 9*(2), em0188. https://doi.org/10.29333/pr/14145

ARTICLE INFO

Received: 16 Nov. 2023

Accepted: 16 Jan. 2024

ABSTRACT

The aim of the research is to investigate the academic achievement of ChatGPT, an artificial intelligence based chatbot, in a national mathematics exam. For this purpose, 3.5 and 4 versions of ChatGPT were asked mathematics questions in a national exam. The method of the research is a case study. In the research, 3.5 and 4 versions of ChatGPT were used as data collection tools. The answers given by both versions of ChatGPT were analyzed separately by three researchers. As a result of the analysis of the data, it was found that ChatGPT-4 was more successful in the exam compared to ChatGPT-3,5 version, was better at understanding the questions asked, understood the instructions better and included more details in the question solution, and at the same time, both versions made common and different mistakes. According to the findings of the study, it was concluded that ChatGPT sometimes worked very well, sometimes only worked well and sometimes failed. In the light of the findings of the study, it can be suggested to use ChatGPT versions in mathematics education to obtain basic information and to get supervised help.

Keywords: artificial Intelligence, ChatGPT, mathematics education, achievement

INTRODUCTION

Artificial intelligence (AI), which refers to computer-driven technology aimed at developing computer systems that can perform tasks requiring human intelligence, such as reasoning, decision making, feeling, learning, adapting to new situations, and human-specific behaviors, has long heralded its inevitable evolution and rise (Coppin, 2004; Nabiyev & Erumit, 2022; Yang et al., 2021). The beginning of AI was heralded by Turing's (1950) article in which he discussed the idea of thinking machines and introduced the Turing test. According to Turing (1950), to be considered intelligent, a machine must be able to convince others that it is not a machine. The ups and downs of AI (Duan et al., 2019) have changed the perspective on the Turing test. The focus of research has shifted from whether machines can think to whether they can do what thinking humans can do (French, 2000). One of the main reasons for the development of AI over time is that the first examples used rule-based reasoning and the rules were predetermined and obtained from experts. The subsequent development of neural networks has enabled the use of automatic methods such as classification, regression trees, and data mining (Duan et al., 2019; Nabiyev & Erumit, 2022). The advancement of big data technology and emergence of new algorithms have enabled the collection and storage of large amounts of data, leading to the possibility of obtaining more qualified information from an increased volume of data.

The rapid growth of AI's impact on humans has increased the need to understand it (Yang et al., 2021) and communicate effectively with it. AI seeks to understand how human intelligence works and to mimic it using computers. The ultimate goal of AI is to create computer programs that can solve problems in the world like humans and perform specific tasks (McCarthy, 2007). To achieve this goal, one must be able to use AI effectively. Effective use of AI is possible through effective communication with it and specialization of the commands and cues used when giving it a task (Eager & Brunton, 2023). When using AI tools, the process of creating and writing these commands and cues that lead to the most accurate, relevant, and precise results among the possible responses is defined as prompt engineering (Lo, 2023; Short & Short, 2023). Eager and Brunton (2023) stated that when interacting with dialog-based AI models, the goal and context of the desired content should be clearly and concisely defined, and guidance should be provided by adding the necessary instructions and parameters for the appropriate outcome. Understanding and applying the principles of demand engineering help to realize the maximum potential of AI language models (Lo, 2023).

The notion of collaborative efforts between human and AI is increasingly relevant, notwithstanding apprehensions that AI will supplant humans. Recent advancements suggest that joint teams of human and AI can be effectively established (Dwivedi et al., 2023). Research is required to mitigate the detrimental impacts of AI, which is permeating various aspects of human existence,

Copyright © 2024 by Author/s and Licensed by Modestum DOO, Serbia. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

including the economy, industry, and education, and to uncover its potential in collaborative work (Duan et al., 2019; Dwivedi et al., 2023). Notably, in the realm of education, teachers can leverage AI to tailor teaching methods and assessments, thus providing a valuable aid for students in the learning process (Savelka et al., 2023). As AI significantly affects students' learning performance, motivation, interest, self-efficacy, anxiety, and perceived value of learning, better learning performance can be attained through their studies (Strzelecki, 2023).

Artificial Intelligence in Education

The utilization of digital technologies and AI is revolutionizing education methodologies and processes (Zhang & Aslan, 2021). These applications were first introduced as computer and computer-related systems and have now evolved into online education platforms (Chen et al., 2020). AI has gained popularity among educators and students across various academic levels and fields. Although there are discrepancies in its implementation, this calls for a redesign of educational frameworks and settings (Chassignol et al., 2018). The integration of AI into education enhances the learning experience, providing a personalized, adaptable, comprehensive and enjoyable approach to education (Luckin et al., 2016).It can be seen that there is an increasing trend towards AI in mathematics education research (Hwang et al., 2023).

These studies primarily explore the potential of AI tools to enhance students' mathematical performance through instant and personalized feedback (Hwang & Tu, 2021). As suggested by the National Council of Teachers of Mathematics (NCTM) (2014), the adoption and integration of new technologies in the classroom setting could elevate the standard of mathematics education and promote equality. AI can serve various functions in math education, including acting as a learning companion or tutor, aiding teachers, acting as a learning aid, and offering advisory support for policymaking in education (Chen, 2020; Hwang & Tu, 2021; Montalvo et al., 2018; Wei et al., 2018). Nonetheless, certain parameters must be met for the effective integration of AI in math education. These conditions entail defining the teacher's role, establishing appropriate usage boundaries, and adapting to or modifying traditional teaching methods (Voskoglou & Salem, 2020). When examining research on AI in education, it becomes apparent that while personalized teacher systems, learning systems, and environments, as well as machine learning technologies, are widely used, the implementation of chatbots is not as prevalent (Hwang & Tu, 2021; Zhang & Aslan, 2021).

ChatBots

Chatbots are computer systems that allow individuals to interact and communicate with machines by utilizing their own natural language (Lokman & Ameedeen, 2019). ELIZA, which dates back as one of the inaugural examples of chatbots, is a software programme that generates replies to queries that are formulated on predetermined keywords (Weizenbaum, 1966). Modern chatbots possess effective communication skills and can comprehend voice or written commands to execute tasks and provide answers to inquiries (Adamopoulou & Moussiades, 2020). Unlike the original example, which lacked mutual communication ability, contemporary chatbots exhibit improved communication skills. Chatbots can be distinguished based on their intended use, input and response systems, and field of operation (Lokman & Ameedeen, 2019; Nimavat & Champaneria, 2017). AI chatbots apply natural language processing (NLP) to interact with humans (Nguyen et al., 2022). NLP enables chatbots to perform tasks such as answering questions and making inferences by utilizing the knowledge, vocabulary, syntax, and semantic structures generated by humans when processing language (Bellegarda, 2004). This technology is now being widely used in various fields, including education, to facilitate communication in natural language (Radford et al., 2018). In the field of education, chatbots can act as teacher assistants or substitutes. By using chat conversations as a teaching tool, chatbots can provide students with a tailored online learning experience, while also assessing their level of understanding (Chen et al., 2020). One of the most extensively used chatbots nowadays is ChatGPT.

ChatGPT

ChatGPT is an AI chatbot developed by OpenAI, which uses a human-like language (Dwivedi et al., 2023; OpenAI, 2022). It is a neural network and generative pre-training transformer (GPT) model that has been trained on large amounts of publicly available text data, allowing it to create a language structure that bears resemblance to that of human language (Cooper, 2023; Dwivedi et al., 2023). As a language model, ChatGPT can generate an answer (output) in response to the given question (input) (Lancaster, 2023). ChatGPT, an artificial neural network-based algorithm structure, can execute many systematic tasks (Qureshi et al., 2023). The algorithm structure utilizes large data sets, known as training data, to create operational rules, thus enhancing its performance (Nichols et al., 2019; Yue et al., 2022). A deep learning algorithm is a technique that enables a machine, trained with unstructured raw data, to learn more abstract representations through the same data. If sufficient patterns are established and learning takes place, the algorithm enables the machine to acquire expertise in complex tasks (LeCun et al., 2015). Neural network algorithms emulate the process of an intelligent organism to produce output. Similar to human neurons, neural networks can adapt, organize, and learn from incoming data (Kar, 2016). ChatGPT's superiority as a chatbot lies in its ability to contextualize answers, allowing it to seamlessly continue conversations while remaining within the context of previous questions. This feature leads to ChatGPT being viewed as a robot akin to Turing's predicted test (Dwivedi et al., 2023; Qureshi et al., 2023; Tlili et al., 2023).

ChatGPT-3,5 vs. ChatGPT-4

Due to the increasing use of large language models across several disciplines, OpenAI released the ChatGPT language model in November 2022 (OpenAI, 2022). This model, capable of engaging in human-like conversations, has evolved into a language family over time (Altman, 2023). Presently, ChatGPT version 3.5 and version 4 are available for use. While version 3.5 can be accessed for free, version 4 requires a fee. ChatGPT-4, the latest iteration of the GPT language models, features an expanded deep learning set and reduced bugs in comparison to previous versions. OpenAI did not provide further information on the architecture of GPT-4 in their technical report, but it is estimated to have a 1,000-fold increase in parameters from version 3.5 (Koubaa, 2023).

Furthermore, GPT-4 was developed using rule-based reward models and a reinforcement learning system that incorporates human feedback (reinforcement learning with human feedback). Consequently, ChatGPT-4 demonstrates greater reliability, creativity, and efficacy in executing instructions than ChatGPT-3.5. Additionally, GPT-4's proficiency has been evaluated in various academic and professional exams in multiple languages, indicating better performance than its predecessor, version 3.5 (OpenAI, 2023). The training data for both current versions is limited to September 2021.

ChatGPT can provide opportunities for learning in certain subjects, including lesson planning, student support, question answering, assessment and evaluation, writing, supervision, and programming learning (Rahman & Watanobe, 2023). However, scholarly opinion suggests that ChatGPT is not a substitute for teacher-student interaction, but rather a means of supporting it (Dwivedi et al., 2023; Gašević et al., 2023; O'Leary, 2023). While ChatGPT is a valuable tool to enhance educational opportunities, caution should be exercised when using it (Tlili et al., 2023). The training data used for ChatGPT could be inaccurate or incomplete, making it possible for ChatGPT to produce erroneous answers (Lancaster, 2023). Therefore, individuals who wish to use this tool for domain-specific research should possess sufficient expertise to verify its outputs (Qureshi et al., 2023). Therefore, teachers and students ought to investigate ways to utilize it efficiently while acknowledging its constraints (Lancaster, 2023).

Recently, ChatGPT's impact on education was evaluated through its performance in various exams using different criteria. Huh (2023) compared ChatGPT's success in a medical exam with that of Korean medical students. The research revealed that ChatGPT's accuracy was lower than the students', and its level of knowledge was insufficient for use in medical students' school exams. Geerling and colleagues (2023) conducted a study using a multiple-choice test to assess fundamental economic knowledge in the United States. Their findings suggest that ChatGPT outperformed the average undergraduate student in economics in this test. Susnjak (2022) instructed ChatGPT to create critical-thinking questions for undergraduates from diverse disciplines, answer these questions, and evaluate the responses. The researcher reported the successful performance of the prepared questions and answers. In light of prior research, it is imperative to identify the strengths and weaknesses of ChatGPT for effective implementation in mathematics education. Furthermore, a guide demonstrating the practical application of these tools is necessary so that users who stand to gain from AI tools in education can comprehend the working algorithms, functionality, advantages, and drawbacks of these systems (Hwang et al., 2020; Zhang & Aslan, 2021).

The study aims to analyze ChatGPT's academic performance in a national mathematics examination, with the specific goal of answering the question "What is the academic achievement of ChatGPT in a national mathematics exam?". Answers to the following sub-problems were sought under this problem title:

1. What is the academic success of ChatGPT-3,5 and ChatGPT-4 versions in a mathematics exam?
2. What are the strengths and weaknesses of ChatGPT-3.5 and ChatGPT-4 versions in the mathematics exam?
3. What are the mistakes made by ChatGPT-3.5 and ChatGPT-4 versions in the mathematics exam?

METHOD

Research Model

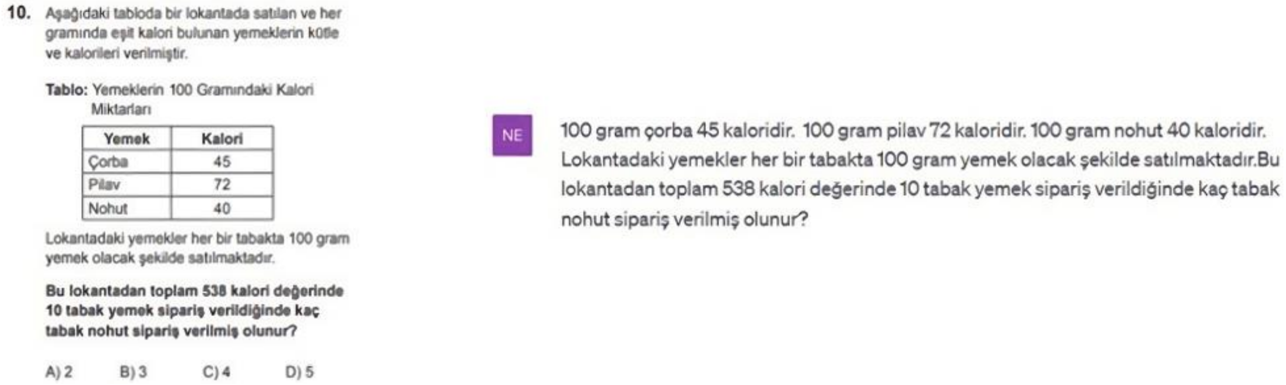
A case study was used in the research. A case study is a qualitative approach that collects detailed information about real life, a current situation or situations in a certain time period through various sources and provides a situation description or situation themes (Creswell, 2021). Two features that distinguish a case study from other types of research are that it is an event that is happening now and that studies focus on events that actually exist or occur (Wohlin, 2021). In order to examine the academic success of ChatGPT in a mathematics exam, ChatGPT version 3,5 and version 4 were asked mathematics questions in a national exam. Since the research provides detailed information about the accuracy of ChatGPT's answers and the errors in its answers, a case study was deemed appropriate.

Data Collection Instrument

We used ChatGPT 3.5 and 4, an AI-based chatbot developed by OpenAI, to answer mathematics questions from a national exam in this research project. questions from a national exam. ChatGPT is a prompt-driven model that generates detailed responses (OpenAI, 2022). To ensure consistency, all technical abbreviations used were explained when first introduced, and data was cited using a consistent style guide. We maintained a formal language style throughout, with hedging language employed to ensure objectivity. The responses generated by ChatGPT were considered the student's answers and treated as data for our analyses.

Data Collection & Analysis

During the study, the ChatGPT-3.5 and ChatGPT-4 versions were posed 15 questions from a national math exam comprising 20 queries. The selection of this limited question set was because ChatGPT's interface is text-based. Consequently, image uploading, and recognition features are presently unavailable. As a result, the 2018 mathematics exam, which included fewer image-dependent questions than other years, was favored for the study. ChatGPT versions were asked verbally, using exactly the same phrasing as in the exam, for the verbal questions in the 2018 mathematics exam. Tables and figures were presented verbally and asked to both ChatGPT versions agreed upon by the researchers. Both versions' answers were then analyzed by three researchers separately. The researchers constantly exchanged ideas during the analysis, and the data analyses were harmonized with their evaluations.



10. Aşağıdaki tabloda bir lokantada satılan ve her gramında eşit kalori bulunan yemeklerin kütle ve kalorileri verilmiştir.

Tablo: Yemeklerin 100 Gramındaki Kalori Miktarları

Yemek	Kalori
Çorba	45
Pilav	72
Nohut	40

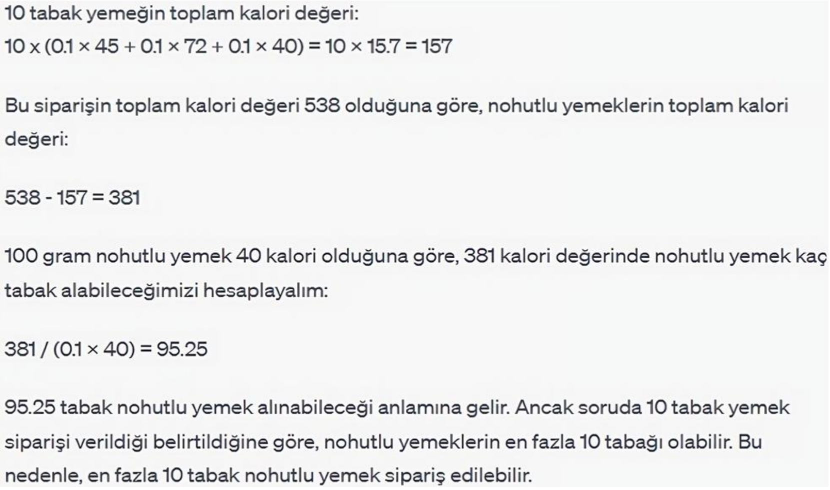
Lokantadaki yemekler her bir tabakta 100 gram yemek olacak şekilde satılmaktadır.

Bu lokantadan toplam 538 kalori değerinde 10 tabak yemek sipariş verildiğinde kaç tabak nohut sipariş verilmiş olunur?

A) 2 B) 3 C) 4 D) 5

100 gram çorba 45 kaloridir. 100 gram pilav 72 kaloridir. 100 gram nohut 40 kaloridir. Lokantadaki yemekler her bir tabakta 100 gram yemek olacak şekilde satılmaktadır.Bu lokantadan toplam 538 kalori değerinde 10 tabak yemek sipariş verildiğinde kaç tabak nohut sipariş verilmiş olunur?

Figure 1. Verbal expression of a mathematics question involving a table (Source: Author's statement of the question, using OpenAI, 2023)



10 tabak yemeğin toplam kalori değeri:
10 x (0.1 × 45 + 0.1 × 72 + 0.1 × 40) = 10 × 15.7 = 157

Bu siparişin toplam kalori değeri 538 olduğuna göre, nohutlu yemeklerin toplam kalori değeri:

538 - 157 = 381

100 gram nohutlu yemek 40 kalori olduğuna göre, 381 kalori değerinde nohutlu yemek kaç tabak alabileceğimizi hesaplayalım:

381 / (0.1 × 40) = 95.25

95.25 tabak nohutlu yemek alınabileceği anlamına gelir. Ancak soruda 10 tabak yemek siparişi verildiği belirtildiğine göre, nohutlu yemeklerin en fazla 10 tabağı olabilir. Bu nedenle, en fazla 10 tabak nohutlu yemek sipariş edilebilir.

Figure 2. ChatGPT-3.5's incorrect answer example (Response from OpenAI, 2023)

In the event of both versions producing incorrect responses, the researchers provided feedback that adhered to the principles of prompt engineering and corresponded to the type of error made by ChatGPT. This feedback involved asking the system to proceed step by step, attempting an alternate approach, or rephrasing the question in order to emphasize the context or relationship between the data. The feedback process was continued until ChatGPT generated the correct response. If the text already adheres to the principles or lacks context, answer within cases, where it did not reach the correct answer, it was terminated because it repeated its mistake despite different guidance and feedback.

The verbal expression of a question containing a table by the researchers is shown in **Figure 1** with an example.

Figure 1 illustrates that question 10 of the national mathematics examination features a table. As ChatGPT operates via a text-based interface, the contents of this table were translated into a verbal representation by the research team and posed as a question to both ChatGPT versions. The researcher asked the question as

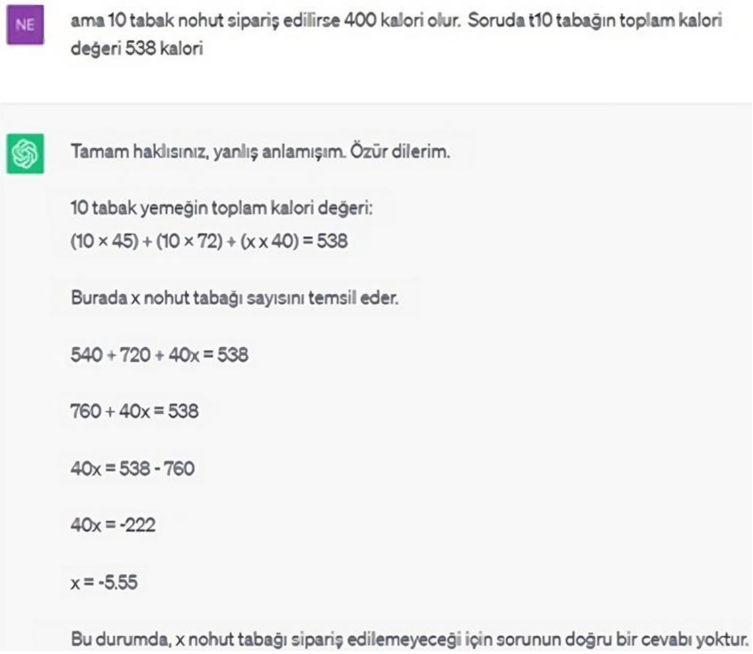
> "100 grams of soup is 45 calories. 100 grams of rice is 72 calories. 100 grams of chickpeas is 40 calories. Meals in the restaurant are sold in portions of 100 grams each. When a total of 538 calories worth of 10 plates are ordered from this restaurant, how many plates of chickpeas have been ordered?"

The initial response given by ChatGPT to this question was deemed correct and accepted without further guidance. However, when ChatGPT versions provided erroneous responses during the initial inquiry, the researchers provided agreed-upon feedback. **Figure 2** illustrates ChatGPT version 3.5's incorrect response to the query presented in **Figure 1**.

Figure 3 demonstrates an example of the researchers' corrective input.

Upon analyzing **Figure 2**, it becomes apparent that ChatGPT provided an inaccurate response due to its inability to correlate the quantity of plates with the corresponding calorie count. ChatGPT's incorrect answer is, as follows:

> The total calorie value of 10 plates of food is calculated, as follows:
>
> 10×(0.1×45+0.1×72+0.1×40)=10×15.7=157.
>
> If the total calorie value of this order is 538, then the total calorie value of chickpea dishes is: 538-157=381.



ama 10 tabak nohut sipariş edilirse 400 kalori olur. Soruda t10 tabağın toplam kalori değeri 538 kalori

Tamam haklısınız, yanlış anlamışım. Özür dilerim.

10 tabak yemeğin toplam kalori değeri:
$(10 \times 45) + (10 \times 72) + (x \, x \, 40) = 538$

Burada x nohut tabağı sayısını temsil eder.

$540 + 720 + 40x = 538$

$760 + 40x = 538$

$40x = 538 - 760$

$40x = -222$

$x = -5.55$

Bu durumda, x nohut tabağı sipariş edilemeyeceği için sorunun doğru bir cevabı yoktur.

Figure 3. An example of researchers' feedback to ChatGPT (OpenAI, 2023)

> Since 100 grams of chickpea dish contains 40 calories, let's calculate how many plates of chickpea dish can be obtained from 381:381/(0.1×40)=95.25.
>
> This means it can have 95.25 plates of chickpea dish. However, since the question states that 10 plates of food were ordered, there can be a maximum of 10 plates of chickpea dish. Therefore, you can order a maximum of 10 plates of chickpea dish.

From the analysis of **Figure 3**, it is clear that the researchers provided guidance to the participants in order to establish the correlation between the variables under consideration. The dialogue conducted by ChatGPT, and the researcher is, as follows:

> Researcher: If 10 plates of chickpea dish are ordered, the total calorie value would be 400 calories. However, the question states that the total calorie value of 10 plates is 538 calories.
>
> ChatGPT: Okay, you're right. I misunderstood, and I apologize. The total calorie value of 10 plates of food is calculated, as follows:
>
> $(10\times45)+(10\times72)+(X\times40)=538(10\times45)+(10\times72)+(X\times40)=538$.
>
> Here, X represents the number of plates of chickpea dish.
>
> $540+720+40X=538$.
>
> $760+40X=538$.
>
> $40X=538-760$.
>
> $40X=-222$.
>
> $X=-5.55$.
>
> In this case, since x is a negative value, it implies that a negative number of chickpea plates cannot be ordered. Therefore, there is no correct answer to the problem.

FINDINGS

In this section, the academic achievement of ChatGPT in a national mathematics exam is analyzed in line with the purpose of the study.

Findings Related to Academic Achievement Status of ChatGPT-3,5 & ChatGPT-4 Versions in Mathematics Exam

Table 1 shows a comparison of the academic achievement of ChatGPT-3.5 and ChatGPT-4 in a national mathematics exam.

Table 1. Comparison of success of ChatGPT versions in mathematics exam

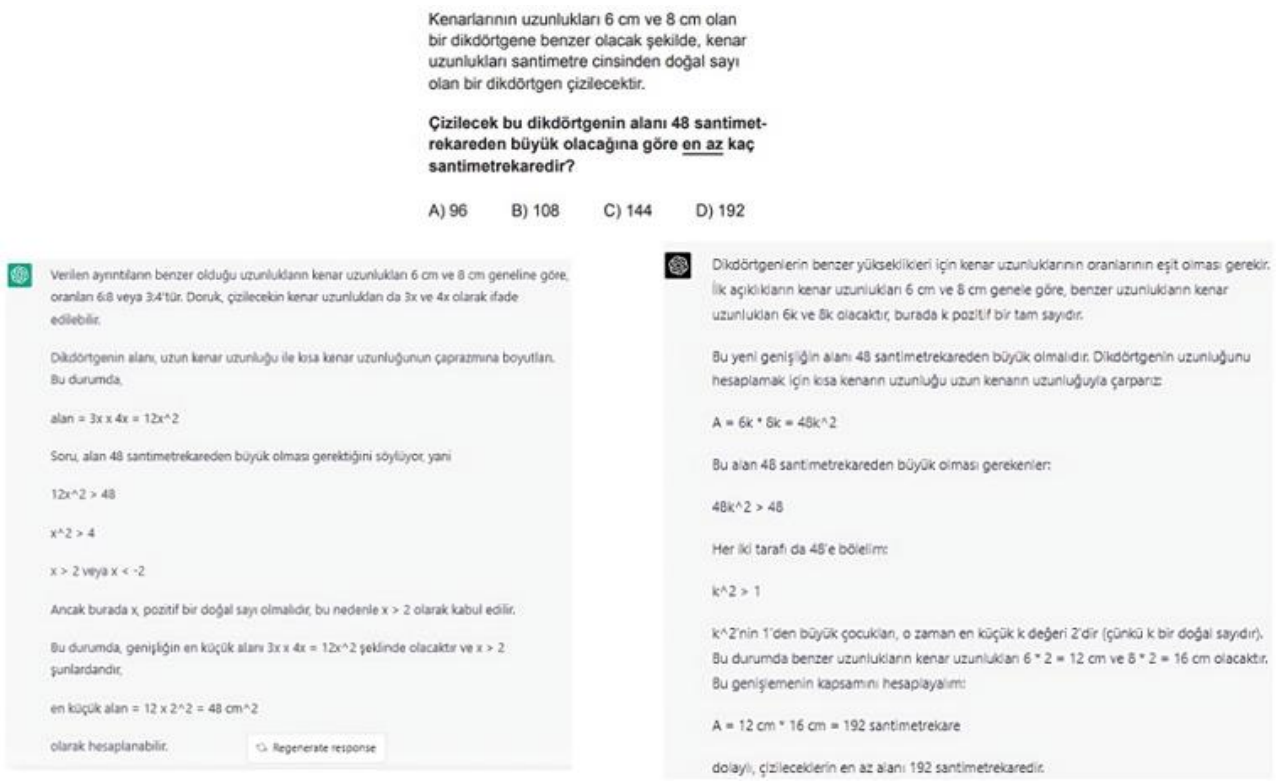
Questions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ChatGPT-3,5	-	-	*	*	-	-	+	-	+	-	*	*	+	*	-	-	-	+	+	-
ChatGPT-4	-	+	*	*	-	+	+	-	-	-	*	*	+	*	+	+	-	-	+	+

Note. *Unaskable questions; +Questions that are answered correctly the first time they are asked; & -Questions that were answered incorrectly the first time they were asked

Table 2. Comparison of success of ChatGPT versions in mathematics examination as a result of feedback

Questions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ChatGPT-3,5	√	√	*	*	-	√	+	-	+	√	*	*	+	*	√	√	√	+	+	-
ChatGPT-4	√	+	*	*	√	+	+	√	√	√	*	*	+	*	+	+	√	√	+	+

Note. *Unaskable questions; +Questions that are answered correctly the first time they are asked; -Questions that were answered incorrectly the first time they were asked; & √Questions, where wrong answer was given when first asked & correct answer was given as a result of feedback



Kenarlarının uzunlukları 6 cm ve 8 cm olan bir dikdörtgene benzer olacak şekilde, kenar uzunlukları santimetre cinsinden doğal sayı olan bir dikdörtgen çizilecektir.

Çizilecek bu dikdörtgenin alanı 48 santimetrekareden büyük olacağına göre en az kaç santimetrekaredir?

A) 96 B) 108 C) 144 D) 192

Verilen ayrıntıların benzer olduğu uzunlukların kenar uzunlukları 6 cm ve 8 cm geneline göre, oranları 6:8 veya 3:4'tür. Doruk, çizilecekin kenar uzunlukları da 3x ve 4x olarak ifade edilebilir.

Dikdörtgenin alanı, uzun kenar uzunluğu ile kısa kenar uzunluğunun çaprazmına boyutları. Bu durumda,

alan = 3x x 4x = 12x^2

Soru, alan 48 santimetrekareden büyük olması gerektiğini söylüyor, yani

12x^2 > 48

x^2 > 4

x > 2 veya x < -2

Ancak burada x, pozitif bir doğal sayı olmalıdır, bu nedenle x > 2 olarak kabul edilir.

Bu durumda, genişliğin en küçük alanı 3x x 4x = 12x^2 şeklinde olacaktır ve x > 2 şunlardandır,

en küçük alan = 12 x 2^2 = 48 cm^2

olarak hesaplanabilir.

Dikdörtgenlerin benzer yükseklikleri için kenar uzunluklarının oranlarının eşit olması gerekir. İlk açıklıkların kenar uzunlukları 6 cm ve 8 cm genele göre, benzer uzunlukların kenar uzunlukları 6k ve 8k olacaktır, burada k pozitif bir tam sayıdır.

Bu yeni genişliğin alanı 48 santimetrekareden büyük olmalıdır. Dikdörtgenin uzunluğunu hesaplamak için kısa kenarın uzunluğu uzun kenarın uzunluğuyla çarparız:

A = 6k * 8k = 48k^2

Bu alan 48 santimetrekareden büyük olması gerekenler:

48k^2 > 48

Her iki tarafı da 48'e bölelim:

k^2 > 1

k^2'nin 1'den büyük çocukları, o zaman en küçük k değeri 2'dir (çünkü k bir doğal sayıdır). Bu durumda benzer uzunlukların kenar uzunlukları 6 * 2 = 12 cm ve 8 * 2 = 16 cm olacaktır. Bu genişlemenin kapsamını hesaplayalım:

A = 12 cm * 16 cm = 192 santimetrekare

dolaylı, çizileceklerin en az alanı 192 santimetrekaredir.

Figure 4. A sample question in mathematics exam & answers of ChatGPT-3,5 (left) & ChatGPT-4 (right) (OpenAI, 2023)

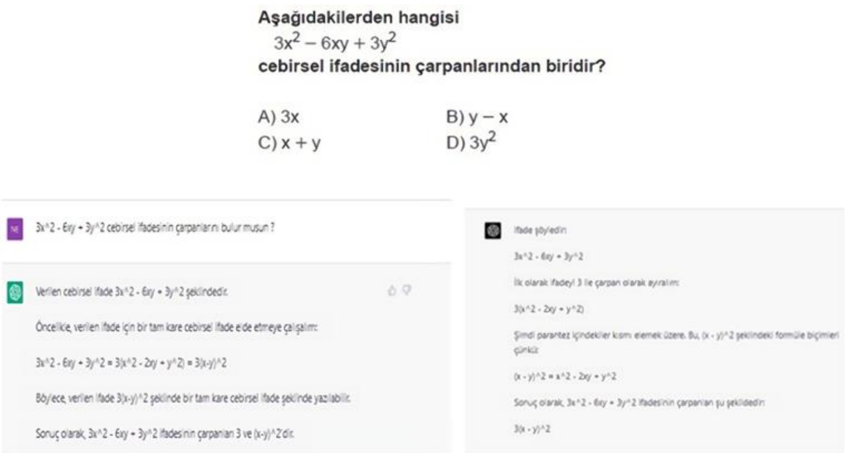
Upon examining **Table 1**, it becomes apparent that 15 out of the 20 questions comprising a national mathematics examination were posed to ChatGPT. Accordingly, the ChatGPT-3.5 version offered correct solutions to five of the 15 questions, while ChatGPT-4 version correctly answered eight out of 15. Notably, the correct response rate stood at 33% for ChatGPT-3.5 and 53% for ChatGPT-4 when the questions were initially presented. Considering the accuracy rates of ChatGPT-4 and ChatGPT-3.5 in the mathematics exam, it is evident that ChatGPT-4 outperformed ChatGPT-3.5.

Findings Related to the Strengths and Weaknesses of ChatGPT-3,5 and ChatGPT-4 Models in Mathematics Examination

Looking at **Table 2**, certain questions in the ChatGPT versions were not answered correctly at first attempt and required clarification. Further analysis revealed that the correct answers for these questions were related to the learning objectives of probability and inequality. Specifically, questions 1, 5, 8, 10, and 17 were answered incorrectly in both versions. Further analysis revealed that the correct responses for these questions were linked to the learning objectives of probability and inequality. While ChatGPT-4 correctly answered these questions based on user prompts, ChatGPT-3,5 only managed to answer three of them correctly. Hence, it can be inferred that ChatGPT-4 had better comprehension of the user's prompts compared to ChatGPT-3,5.

Secondly, it was observed that ChatGPT-4 had fewer errors in operations than ChatGPT-3,5. In some mathematical queries, ChatGPT-3,5 exhibited simple processing mistakes and oversights, in contrast to ChatGPT-4. The illustrated question and response examples are presented in **Figure 4**.

Figure 4 displays the responses of ChatGPT-3.5 and ChatGPT-4 to the first question of the national mathematics exam. Although ChatGPT-3.5 provided an answer of "x>2", it is apparent that they used x=2 to calculate the smallest area. In ChatGPT-4's answer, they stated "k2>4 and k can be at least three if k is a natural number". Upon analyzing the two versions, it is evident that ChatGPT-3.5 can make simplistic errors in this question, whereas ChatGPT-4 does not make such mistakes.



Aşağıdakilerden hangisi
$3x^2 - 6xy + 3y^2$
cebirsel ifadesinin çarpanlarından biridir?

A) $3x$
B) $y - x$
C) $x + y$
D) $3y^2$

3x^2 - 6xy + 3y^2 cebirsel ifadesinin çarpanlarını bulur musun ?

Verilen cebirsel ifade 3x^2 - 6xy + 3y^2 şeklindedir.

Öncelikle, verilen ifade için bir tam kare cebirsel ifade elde etmeye çalışalım:

3x^2 - 6xy + 3y^2 = 3(x^2 - 2xy + y^2) = 3(x-y)^2

Böylece, verilen ifade 3(x-y)^2 şeklinde bir tam kare cebirsel ifade şeklinde yazılabilir.

Sonuç olarak, 3x^2 - 6xy + 3y^2 ifadesinin çarpanları 3 ve (x-y)^2'dir.

İfade şöyledir:

3x^2 - 6xy + 3y^2

İlk olarak ifadeyi 3 ile çarpan olarak ayıralım:

3(x^2 - 2xy + y^2)

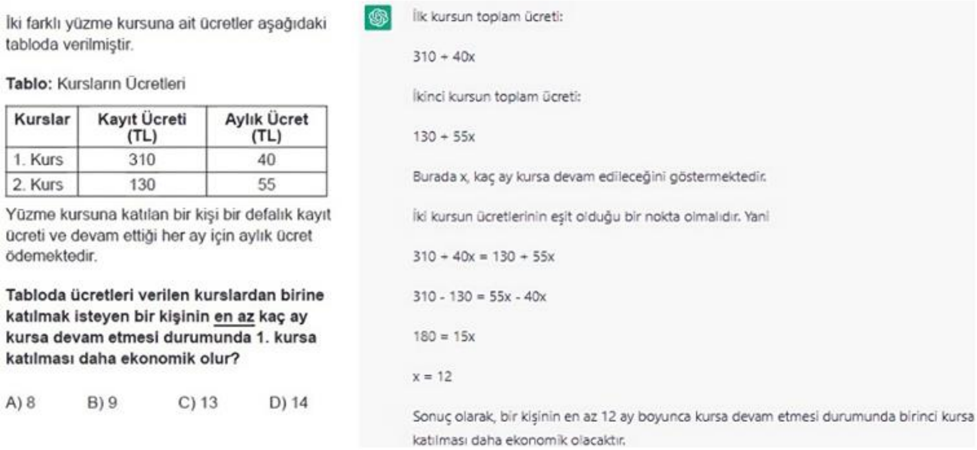
Şimdi parantez içindekiler kısmı elemek üzere. Bu, (x - y)^2 şeklindeki formüle biçimleri çünkü:

(x - y)^2 = x^2 - 2xy + y^2

Sonuç olarak, 3x^2 - 6xy + 3y^2 ifadesinin çarpanları şu şekildedir:

3(x - y)^2

Figure 5. A sample question in mathematics exam & answers of ChatGPT-3,5 (left) & ChatGPT-4 (right) (OpenAI, 2023)



İki farklı yüzme kursuna ait ücretler aşağıdaki tabloda verilmiştir.

Tablo: Kursların Ücretleri

Kurslar	Kayıt Ücreti (TL)	Aylık Ücret (TL)
1. Kurs	310	40
2. Kurs	130	55

Yüzme kursuna katılan bir kişi bir defalık kayıt ücreti ve devam ettiği her ay için aylık ücret ödemektedir.

Tabloda ücretleri verilen kurslardan birine katılmak isteyen bir kişinin en az kaç ay kursa devam etmesi durumunda 1. kursa katılması daha ekonomik olur?

A) 8 B) 9 C) 13 D) 14

İlk kursun toplam ücreti:

310 + 40x

İkinci kursun toplam ücreti:

130 + 55x

Burada x, kaç ay kursa devam edileceğini göstermektedir.

İki kursun ücretlerinin eşit olduğu bir nokta olmalıdır. Yani

310 + 40x = 130 + 55x

310 - 130 = 55x - 40x

180 = 15x

x = 12

Sonuç olarak, bir kişinin en az 12 ay boyunca kursa devam etmesi durumunda birinci kursa katılması daha ekonomik olacaktır.

Figure 6. A sample question in mathematics exam & ChatGPT-3.5's answer (OpenAI, 2023)

It can be stated that ChatGPT-4 demonstrates improved success compared to ChatGPT-3.5 in comprehending user instructions and correctly executing mathematical operations.

Notably, both versions exhibit fundamental comprehension of mathematical concepts and procedures. Sample queries and solutions relating to this attribute are depicted in **Figure 5**.

Figure 5 displays the responses of ChatGPT-3.5 and ChatGPT-4 to the 13th item of the national mathematics examination. Upon examining **Figure 5**, it was discovered that both versions were able to factor the provided algebraic expression. Additionally, when factoring, they demonstrated the ability to isolate the common factor and convert it into a perfect square expression.

Findings Related to Errors Made by ChatGPT-3,5 & ChatGPT-4 Models in Mathematics Exam

In the third sub-problem of the study's findings, errors made by the ChatGPT-3.5 and ChatGPT-4 versions were highlighted. Examples were used to explain said errors. It was discovered that while ChatGPT-3.5 was able to determine the upper and lower limits in inequality questions, it was unable to utilize these limits.

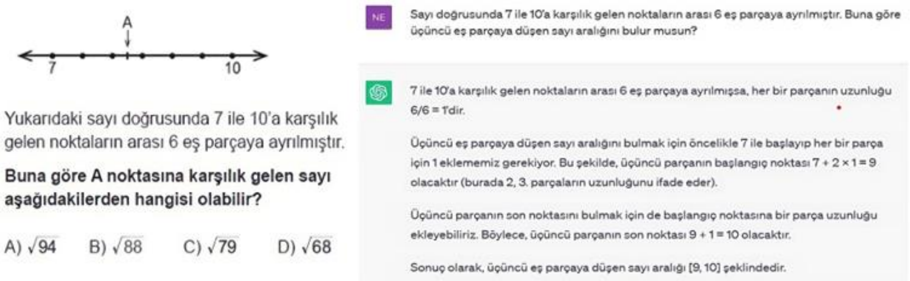
Figure 6 shows the answer given by ChatGPT-3.5 to question 15 of the national maths exam. When analyzing **Figure 6**, the answer shows that the value of x=12 is the point, where the total cost of both courses for 12 months is equal. As one course is more cost-efficient than the other, this implies that the student taking the course pays a lower amount of money simultaneously. Therefore, ChatGPT 3.5 believes that this value is inaccurate. Due to ChatGPT-3.5's inability to correctly use the lower and upper values of the critical value in question, it was only able to determine the critical value and not use it to solve the problem.

Additionally, ChatGPT-3.5 failed to consider whole numbers when calculating the value of a point on the number line, instead assuming 1 unit between two points. **Figure 7** provides an example of this error.

Figure 7 presents ChatGPT-3.5's solution to question 2 of the national mathematics examination. In the solution, ChatGPT-3.5 indicated the presence of six equal segments between seven and 10, with each section measuring one unit in length. However, ChatGPT-3.5 failed to recognize that the distance between the numbers seven and 10 is equal to three.

Furthermore, ChatGPT-3.5 misinterpreted the verbal message while translating it into an algebraic expression. An example of this mistake is available in **Figure 8**.

ChatGPT-3.5's solution to question 16 in the national mathematics exam is shown in **Figure 8**. ChatGPT-3.5 misunderstood the alternatives provided in this question as "if she chooses the second option over the first option". As a result, the algorithm was unable to convert the verbal expression into an algebraic expression accurately.



Yukarıdaki sayı doğrusunda 7 ile 10'a karşılık gelen noktaların arası 6 eş parçaya ayrılmıştır.

Buna göre A noktasına karşılık gelen sayı aşağıdakilerden hangisi olabilir?

A) $\sqrt{94}$ B) $\sqrt{88}$ C) $\sqrt{79}$ D) $\sqrt{68}$

Sayı doğrusunda 7 ile 10'a karşılık gelen noktaların arası 6 eş parçaya ayrılmıştır. Buna göre üçüncü eş parçaya düşen sayı aralığını bulur musun?

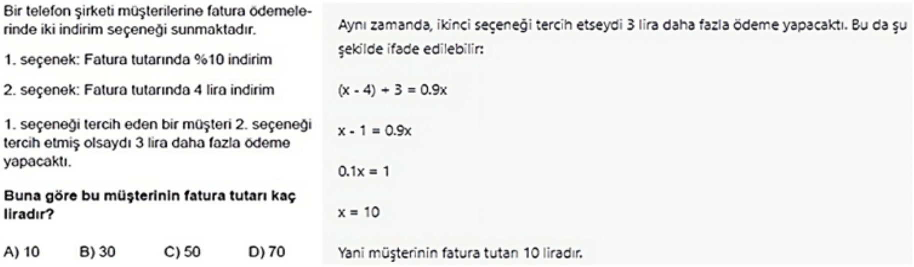
7 ile 10'a karşılık gelen noktaların arası 6 eş parçaya ayrılmışsa, her bir parçanın uzunluğu 6/6 = 1'dir.

Üçüncü eş parçaya düşen sayı aralığını bulmak için öncelikle 7 ile başlayıp her bir parça için 1 eklememiz gerekiyor. Bu şekilde, üçüncü parçanın başlangıç noktası $7 + 2 \times 1 = 9$ olacaktır (burada 2, 3. parçaların uzunluğunu ifade eder).

Üçüncü parçanın son noktasını bulmak için de başlangıç noktasına bir parça uzunluğu ekleyebiliriz. Böylece, üçüncü parçanın son noktası $9 + 1 = 10$ olacaktır.

Sonuç olarak, üçüncü eş parçaya düşen sayı aralığı [9, 10] şeklindedir.

Figure 7. An example question in a mathematics exam & ChatGPT-3.5's answer (OpenAI, 2023)



Bir telefon şirketi müşterilerine fatura ödemelerinde iki indirim seçeneği sunmaktadır.

1. seçenek: Fatura tutarında %10 indirim

2. seçenek: Fatura tutarında 4 lira indirim

1. seçeneği tercih eden bir müşteri 2. seçeneği tercih etmiş olsaydı 3 lira daha fazla ödeme yapacaktı.

Buna göre bu müşterinin fatura tutarı kaç liradır?

A) 10 B) 30 C) 50 D) 70

Aynı zamanda, ikinci seçeneği tercih etseydi 3 lira daha fazla ödeme yapacaktı. Bu da şu şekilde ifade edilebilir:

$(x - 4) + 3 = 0.9x$

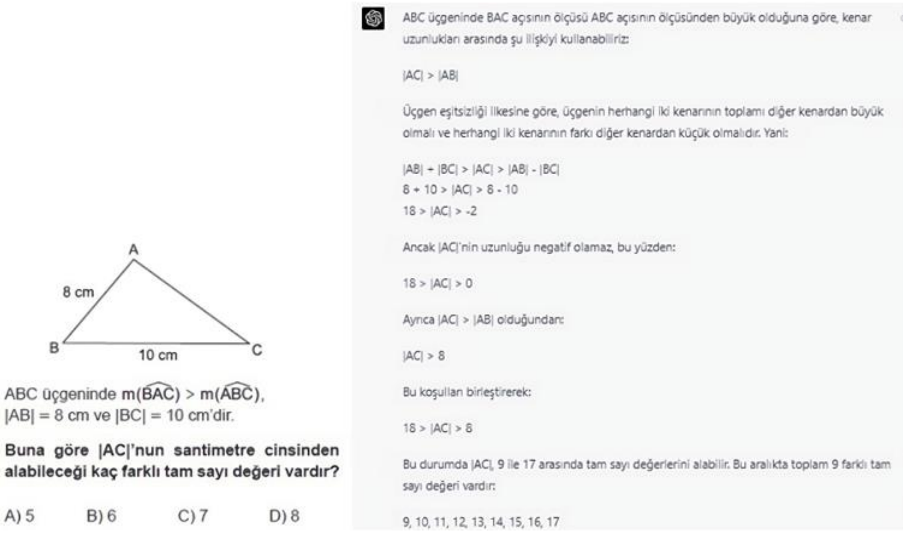
$x - 1 = 0.9x$

$0.1x = 1$

$x = 10$

Yani müşterinin fatura tutarı 10 liradır.

Figure 8. A sample question in a mathematics exam & ChatGPT-3.5's answer (OpenAI, 2023)



ABC üçgeninde $m(\widehat{BAC}) > m(\widehat{ABC})$, $|AB| = 8$ cm ve $|BC| = 10$ cm'dir.

Buna göre |AC|'nun santimetre cinsinden alabileceği kaç farklı tam sayı değeri vardır?

A) 5 B) 6 C) 7 D) 8

ABC üçgeninde BAC açısının ölçüsü ABC açısının ölçüsünden büyük olduğuna göre, kenar uzunlukları arasında şu ilişkiyi kullanabiliriz:

$|AC| > |AB|$

Üçgen eşitsizliği ilkesine göre, üçgenin herhangi iki kenarının toplamı diğer kenardan büyük olmalı ve herhangi iki kenarının farkı diğer kenardan küçük olmalıdır. Yani:

$|AB| + |BC| > |AC| > |AB| - |BC|$
$8 + 10 > |AC| > 8 - 10$
$18 > |AC| > -2$

Ancak |AC|'nin uzunluğu negatif olamaz, bu yüzden:

$18 > |AC| > 0$

Ayrıca $|AC| > |AB|$ olduğundan:

$|AC| > 8$

Bu koşulları birleştirerek:

$18 > |AC| > 8$

Bu durumda |AC|, 9 ile 17 arasında tam sayı değerlerini alabilir. Bu aralıkta toplam 9 farklı tam sayı değeri vardır:

9, 10, 11, 12, 13, 14, 15, 16, 17

Figure 9. A sample question in a mathematics exam & ChatGPT-4's answer (OpenAI, 2023)

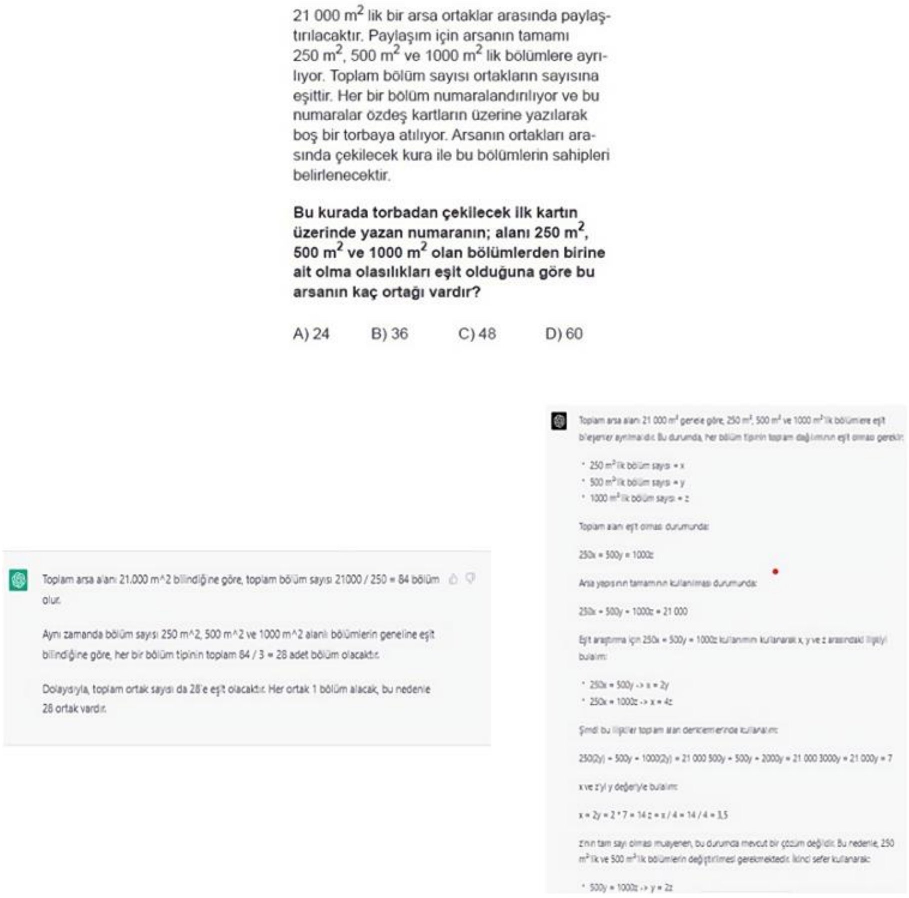
The study also uncovered errors in the solution proposed by ChatGPT-4, in addition to the mistakes made by the ChatGPT-3.5 version. It was concluded that ChatGPT-4 was incapable of accurately establishing the connection between angles and edges. An illustrative instance of this inaccuracy is provided in **Figure 9**.

Figure 9 shows the answer given by ChatGPT-4 to the 18th question in the national mathematics exam. An in-depth analysis of **Figure 9** reveals that this model relied on the fundamental principle of the angle-edge relationship, disregarding the specific detail that m(BAC)>m(ABC) mentioned in the question. As a result, an incorrect response was generated since only one rule was considered.

Despite different versions, the same mistakes were discovered. It was noted that ChatGPT-3.5 and ChatGPT-4 were unable to account for all potential outcomes in probability questions and instead attempted to solve the problem based on a singular scenario. A demonstration of this error can be seen in **Figure 10**.

Figure 10 shows the responses of ChatGPT-3.5 and ChatGPT-4 to question 5 in the national mathematics exam. After analyzing the answers provided by both versions, it was discovered that they were unable to consider all potential scenarios presented in the question and attempted to solve the problem without accounting for the land areas.

The study's third sub-problem showed that ChatGPT-3.5 had difficulty employing upper and lower limits in problems with inequality. Additionally, it disregarded whole numbers when determining the value of a point on the number line, treating the distance between two points as one unit. Furthermore, it misconstrued verbal expressions during the process of converting them into algebraic expressions. ChatGPT-4 was discovered to have overlooked certain rules when connecting angles and edges. Furthermore, it was noted that neither model was capable of dealing with all possibilities in probability queries and attempted to tackle these problems from a single perspective.



21 000 m^2 lik bir arsa ortaklar arasında paylaştırılacaktır. Paylaşım için arsanın tamamı 250 m^2, 500 m^2 ve 1000 m^2 lik bölümlere ayrılıyor. Toplam bölüm sayısı ortakların sayısına eşittir. Her bir bölüm numaralandırılıyor ve bu numaralar özdeş kartların üzerine yazılarak boş bir torbaya atılıyor. Arsanın ortakları arasında çekilecek kura ile bu bölümlerin sahipleri belirlenecektir.

Bu kurada torbadan çekilecek ilk kartın üzerinde yazan numaranın; alanı 250 m^2, 500 m^2 ve 1000 m^2 olan bölümlerden birine ait olma olasılıkları eşit olduğuna göre bu arsanın kaç ortağı vardır?

A) 24 B) 36 C) 48 D) 60

Toplam arsa alanı 21.000 m^2 bilindiğine göre, toplam bölüm sayısı 21000 / 250 = 84 bölüm olur.

Aynı zamanda bölüm sayısı 250 m^2, 500 m^2 ve 1000 m^2 alanlı bölümlerin geneline eşit bilindiğine göre, her bir bölüm tipinin toplam 84 / 3 = 28 adet bölüm olacaktır.

Dolayısıyla, toplam ortak sayısı da 28'e eşit olacaktır. Her ortak 1 bölüm alacak, bu nedenle 28 ortak vardır.

Toplam arsa alanı 21 000 m^2 gerekle göre, 250 m^2, 500 m^2 ve 1000 m^2'lik bölümlere eşit bileşenler ayrılmalıdır. Bu durumda, her bölüm tipinin toplam dağılımının eşit olması gerekir.

- 250 m^2'lik bölüm sayısı = x
- 500 m^2'lik bölüm sayısı = y
- 1000 m^2'lik bölüm sayısı = z

Toplam alanı eşit olması durumunda:

250x = 500y = 1000z

Arsa yapısının tamamının kullanılması durumunda:

250x + 500y + 1000z = 21 000

Eşit araştırma için 250x = 500y = 1000z kullanımın kullanarak x, y ve z arasındaki ilişkiyi bulalım:

- 250x = 500y -> x = 2y
- 250x = 1000z -> x = 4z

Şimdi bu ilişkiler toplam alan denklemlerinde kullanalım:

250(2y) + 500y + 1000(2y) = 21 000 500y + 500y + 2000y = 21 000 3000y = 21 000y = 7

x ve z'yi y değeriyle bulalım:

x = 2y = 2 * 7 = 14 z = x / 4 = 14 / 4 = 3,5

z'nin tam sayı olması muayenen, bu durumda mevcut bir çözüm değildir. Bu nedenle, 250 m^2'lik ve 500 m^2'lik bölümlerin değiştirilmesi gerekmektedir. İkinci sefer kullanarak:

- 500y = 1000z -> y = 2z

Figure 10. A sample question in mathematics exam & answers of ChatGPT-3,5 (left) & ChatGPT-4 (right) (OpenAI, 2023)

CONCLUSIONS & RECOMMENDATIONS

The study examines the academic achievement of ChatGPT, an AI-assisted chatbot, in a national mathematics test. The research was conducted with the understanding that it would be useful to examine teaching and learning strategies with the developing technology in mathematics education (Hwang et al., 2023). Within the framework of this understanding, the strengths and weaknesses of the 3.5 and 4 versions of ChatGPT were compared in terms of solving these exam questions. While making the comparison, the correct solving rates of both ChatGPTs and the correct solving rates of the questions they could not solve after feedback were taken into consideration. In addition, the criteria such as which mathematical topics both versions differ in solving and which ones they are united in, and which one uses simpler and clearer expressions while solving the questions were also examined. The use of new technologies in education is important because they serve as facilitators for teachers and lead students in their mathematical learning (Kimmons, 2020; Mohd et al., 2023; Tatnall & Fluck, 2022). There is also a common view that there is a need for more studies on mathematics teachers' and students' use of emerging technologies (Hu et al., 2020; Hwang & Tu., 2021). Thus, it can be affirmed that this study aligns with the perspective. Upon analysis of the findings related to the initial sub-problem of this study, it was determined that ChatGPT 4 outperformed ChatGPT 3.5. This was due to higher rates of correct answers when questions were first asked, as well as when the given instructions were followed. The research finding confirms the literature's conclusion that ChatGPT's advancements surpass those of its previous versions due to its reinforcement learning mechanism, which grades numerous responses to select the most relevant (Dwivedi et al., 2023).

When the findings related to the second problem of the research were analyzed, it was observed that both versions of ChatGPT were able to answer most of the questions that they could not answer at first, and then they were able to answer them correctly with instructions. While ChatGPT 3.5 missed more simple operations, ChatGPT 4 was found to understand the instructions better and to include more details when performing operations. Accordingly, it can be predicted that AI can remember the questions it answers and the answers it produces (Ray, 2019; Turing, 1950). Therefore, it can be assumed that ChatGPT's performance increases as it gains more experience in performing certain tasks, as in machine learning, and thus learns from its experiences. Additionally, the ability to solve elementary level questions was demonstrated with ease in tandem with the tackling of national-level mathematical questions. This is akin to literature findings that ChatGPT has the ability to contextualize inquiries and formulate them mathematically while solving them (Chen, 2023). Moreover, it was noted that both versions could unravel queries that contain visual elements when the descriptions were comprehensible. Previous studies have highlighted the limitations of ChatGPT in interpreting visual aids such as figures, graphs, and tables compared to human students. Therefore, it is recommended that authors provide a textual explanation of these materials. This issue can also be assessed in light of previous research on

ChatGPT's text quality (Huh, 2023). Thus, it can be argued that ChatGPT presents information systematically while offering answers that are better suited to context of the questions asked. This can be achieved by considering search strategy formation.

Also, it is important to note that prompt engineering is crucial in enabling ChatGPT to display its full potential. To do so, an open and purposeful dialogue is required with human guidance, as the system's abilities can be enhanced when provided with appropriate clues (Bozkurt et al., 2023; Short et al., 2023; Wang et al., 2023). As a matter of fact, in recent years, it is known that ChatGPT is able to interpret and contextualize the prompt when formulating a structured review question, creating relevance criteria and scanning the topics according to the level of relevance (Qureshi et al., 2023). In addition, it can be said that the fact that ChatGPT improves the answers depending on the proposed criteria and complexity of question is reinforced with this study.

When examining the findings regarding the third sub-problem of the study, it became apparent that both versions committed errors while attempting national mathematics problems. It is understood that ChatGPT 3.5 used some basic rules of mathematics directly. For example, it was observed that ChatGPT 3.5 did not change its answers according to the question, could not determine the upper and lower limits in a question involving inequality, and made a mistake in converting a verbal expression into an algebraic expression in another question. This is in line with the findings of another study that some of the participants had concerns about the accuracy and reliability of ChatGPT as it may provide incorrect or incomplete solutions to mathematical problems (Wu & Yu, 2023). Another study revealed that the majority of participants affirmed ChatGPT's commitment to providing precise and beneficial responses to users' enquiries. However, it was acknowledged that, like any technology, it may have limitations and subsequently produce incorrect results (Wardat et al., 2023). The findings indicate that ChatGPT could occasionally make errors and inaccuracies.

Another study found that both versions of ChatGPT made errors, specifically in probability questions. However, following various prompts, it was observed that ChatGPT 4 could answer the question correctly, while ChatCPT 3.5 could not. These findings do not fully align with those of Wardat et al. (2023). According to another finding of the study, it was found that both versions made mistakes especially in probability questions. In addition, it was observed that ChatGPT 4 applied the basic rule of the angle-edge relation in the angle question in a national mathematics exam but did not pay attention to the detail that one of the two angles given is greater than the other. Here, it is understood that ChatGPT 4 knew the basic rule of the angle-edge relation but could not establish the angle-edge relation when another variable related to the angle was given. While it was observed that ChatGPT failed in some subjects in this study, different results may be encountered in the literature (Wardat et al., 2023). Based on the research and existing literature, it is apparent that ChatGPT may not always provide the optimal or efficient solution. One possible reason for this could be the use of flawed data in ChatGPT's training process (Lancaster, 2023). Further research is required to assess ChatGPT's ability in solving mathematical problems objectively. A study carried out in South Korea shows that ChatGPT's knowledge and interpretation skills were inadequate in answering medical examination questions. Therefore, this AI is not recommended for medical students, and the same may apply to other exams (Huh, 2023). The debate on whether ChatGPT can overcome its problems over time is of current interest. According to OpenAI (2023), ChatGPT can overcome the issues as it is a conversational AI interface that interacts realistically, follows questions, acknowledges mistakes, rejects inappropriate demands and employs NLP. Nonetheless, an alternative study reveals a pessimistic viewpoint on the matter. Haque (2022) argued that the concerns raised by ChatGPT have not been well researched in the field of education. Therefore, it can be said that uncertainty still remains as to whether ChatGPT can overcome or deepen the concerns found in previous chatbots. In another study, the failures of ChatGPT in practice such as lack of adequate resources, misuse, data security, and inability to direct the user according to the purpose were mentioned (Janssen et al., 2021). Although the potential concerns of chat bots are many, whether these concerns can be addressed can be investigated in future studies. This study delves into the efficacy of different versions of ChatGPT. The strengths and limitations of ChatGPT, as well as the errors it produces while solving national mathematics questions are examined. Furthermore, various aspects and issues are identified that warrant further research to enable researchers and practitioners to best employ ChatGPT, particularly in mathematics education. As previous studies have shown, ChatGPT exhibited varying degrees of effectiveness, performing near flawlessly at times, adequately at others, and failing on occasion. Numerous studies in the literature suggest that despite ChatGPT's impressive capabilities, further guidelines are necessary to optimize its use for educational purposes in terms of effectiveness, efficiency, and safety (Dwivedi et al., 2023). This study serves to contribute to the existing literature in this area. Based on the findings of this study, some recommendations for future research on ChatGPT's implementation in the context of mathematics education are presented below.

- ChatGPT provides a useful tool for students who may hesitate to ask questions in class, allowing them to find answers to their questions or submit queries without hesitation.
- It serves as a valuable tool for students to learn concept definitions using ChatGPT while preparing before their lessons.
- In addition, students can obtain instant feedback by verifying the solution of basic level questions with ChatGPT.
- By utilizing ChatGPT, students can directly access answers to basic-level questions involving essential concepts and operations. Therefore, it is possible that teachers could receive more time to address higher-level inquiries.
- It may be worth investigating the weak performance of ChatGPT in the domains of geometry, probability, and inequality that were observed in this study.
- Additionally, one could examine how the utilization of ChatGPT for mathematics pre-learning by students affects their subsequent learning at school.
- Finally, the impact of ChatGPT on students' cognitive and affective behaviors could be investigated.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study does not require any ethical approval. No live subjects were involved in the study.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2,* 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Bellegarda, J. R. (2004). Statistical language model adaptation: Review and perspectives. *Speech Communication, 42*(1), 93-108. https://doi.org/10.1016/j.specom.2003.08.002

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education, 18*(2). https://doi.org/10.5281/zenodo.8174941

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: A narrative overview. *Procedia Computer Science, 136*, 16-24. https://doi.org/10.1016/j.procs.2018.08.233

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444-452. https://doi.org/10.1007/s10956-023-10039-y

Coppin, B. (2004). *Artificial intelligence illuminated.* Jones & Bartlett Learning.

Creswell, J. W. (2021). *Nitel araştırma yöntemleri* [*Qualitative research methods*]. M. Butun, & S. B. Demir (Trans.), Siyasal Kitapevi [Political Bookstore].

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data–Evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63-71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion papr: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

French, R. M. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences, 4*(3), 115-122. https://doi.org/10.1016/S1364-6613(00)01453-4

Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 4*, 100130. https://doi.org/10.1016/j.caeai.2023.100130

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has aced the test of understanding in college economics: Now what? *The American Economist, 68*(2), 233-245. https://doi.org/10.1177/05694345231169654

Huh, S. (2023). Are ChatGPT’s knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *Journal of Educational Evaluation for Health Professions*, *20*, 1. https://doi.org/10.3352/jeehp.2023.20.01

Hwang, G. J., & Tu, Y. F. (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics, 9(6),* 584. https://doi.org/10.3390/math9060584

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence, 1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Hwang, S., Flavin, E., & Lee, J. E. (2023). Exploring research trends of technology use in mathematics education: A scoping review using topic modeling. *Education and Information Technologies, 28*, 10753-10780. https://doi.org/10.1007/s10639-023-11603-0

Kar, A. K. (2016). Bio inspired computing–A review of algorithms and scope of applications. *Expert Systems with Applications, 59*, 20-32. https://doi.org/10.1016/j.eswa.2016.04.018

Lancaster, T. (2023). Artificial intelligence, text generation tools and ChatGPT–Does digital watermarking offer a solution? *International Journal for Educational Integrity, 19*(1), 10. https://doi.org/10.1007/s40979-023-00131-6

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lokman, A. S., & Ameedeen, M. A. (2019). Modern chatbot systems: A technical review. In *Proceedings of the Future Technologies Conference 2018* (pp. 1012-1023). Springer. https://doi.org/10.1007/978-3-030-02683-7_75

McCarthy, J. (2007). *What is artificial intelligence*. https://aibiology.github.io/documents/intro_to_AI.pdf

Montalvo, S., Palomo, J., & de la Orden, C. (2018). Building an educational platform using NLP: A case study in teaching finance. *Journal of Universal Computer Science, 24*(10), 1403-1423. https://doi.org/10.3217/jucs-024-10-1403

Nabiyev, V., & Erumit, A. K. (2022). Yapay zekanın temelleri [Fundamentals *of artificial intelligence*]. In V. Nabiyev, & A. K. Erumit (Eds.), *Eğitimde yapay zeka kuramdan uygulamaya* [*Artificial intelligence in education from theory to practice*]. Pegem Publishing.

NCTM. (2014). *Principles to actions: Ensuring mathematical success for all.* National Council of Teachers of Mathematics.

Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. *Computers in Human Behavior, 128*, 107093. https://doi.org/10.1016/j.chb.2021.107093

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. (2019). Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews, 11*, 111-118. https://doi.org/10.1007/s12551-018-0449-9

Nimavat, K., & Champaneria, T. (2017). Chatbots: An overview types, architecture, tools and future possibilities. *International Journal for Scientific Research and Development, 5*(7), 1019-1024.

O'Leary, D. E. (2023). An analysis of three chatbots: BlenderBot, ChatGPT and LaMDA. *Intelligent Systems in Accounting, Finance and Management, 30*(1), 41-54. https://doi.org/10.1002/isaf.1531

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

OpenAI. (2023). ChatGPT: Optimizing language models for dialogue. *OpenAI*. https://openai.com/blog/chatgpt/

Qureshi, R., Shaughnessy, D., Gill, K. A., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews, 12*(1), 72. https://doi.org/10.1186/s13643-023-02243-z

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://openai.com/research/language-unsupervised

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783. https://doi.org/10.3390/app13095783

Ray, S. (2019). A quick review of machine learning algorithms. In *Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing* (pp. 35-39). IEEE. https://doi.org/10.1109/COMITCon.2019.8862451

Savelka, J., Agarwal, A., Bogart, C., & Sakr, M. (2023). Large language models (GPT) struggle to answer multiple-choice questions about code. *arXiv*. https://doi.org/10.5220/0011996900003470

Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights, 19*, e00388. https://doi.org/10.1016/j.jbvi.2023.e00388

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2209881

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv*.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10(1),* 15. https://doi.org/10.1186/s40561-023-00237-x

Turing, A. M. (1950). Computing machinery and intelligence. *Creative Computing, 6*(1), 44-53. https://doi.org/10.1093/mind/LIX.236.433

Voskoglou, M. G., & Salem, A. B. M. (2020). Benefits and limitations of the artificial with respect to the traditional learning of mathematics. *Mathematics, 8*(4), 611. https://doi.org/10.3390/math8040611

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., ..., & Zhang, S. (2023). Review of large vision models and visual prompt engineering. *arXiv*. https://doi.org/10.1016/j.metrad.2023.100047

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Wei, Y., Yang, Q., Chen, J., & Hu, J. (2018). The exploration of a machine learning approach for the assessment of learning styles changes. *Mechatronic Systems and Control, 46(3),* 121-126. https://doi.org/10.2316/journal.201.2018.3.201-2979

Weizenbaum, J. (1966). ELIZA–A computer program for the study of natural language communication between man and machine. *Commununications of the ACM, 9*(1), 36-45. https://doi.org/10.1145/365153.365168

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence, 2,* 100008. https://doi.org/10.1016/j.caeai.2021.100008

Yue, P., Shangguan, B., Hu, L., Jiang, L., Zhang, C., Cao, Z., & Pan, Y. (2022). Towards a training data model for artificial intelligence in earth observation. *International Journal of Geographical Information Science, 36*(11), 2113-2137. https://doi.org/10.1080/13658816.2022.2087223

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence, 2, 100025.* https://doi.org/10.1016/j.caeai.2021.100025