Performance of Generative Artificial Intelligence Models (GPT-4o, Gemini, Copilot) in YKS/TYT Exam: A Comparative Study

Araştırma Makalesi/Research Article

Selma BULUT[1]

[1]Department of Computer Technologies, Kırklareli University, Kırklareli, Türkiye

selma.bulut@klu.edu.tr

(Geliş/Received:30.10.2024; Kabul/Accepted:31.07.2025)
DOI: 10.17671/gazibtd.1575755

***Abstract*—** This study attempts to evaluate the potential of generative artificial intelligence (GAI) models in the field of education. A simulated exam environment was planned and error analysis was performed based on the responses of GPT (Generative Pre-trained Transformer) family models to questions asked in university entrance exams. In the study, GPT, Gemini and Copilot models were tested with questions in the Higher Education Institutions Exam (YKS)/ Basic Proficiency Exam (TYT) session in Turkey. The responses produced by these models were analyzed in terms of both accuracy rates and academic depth and consistency. While GPT-4o achieved 50.8% accuracy, Gemini 43.3% and Copilot achieved 42.5% accuracy. The findings provide important clues about the potential of GAI models to be used as a learning aid in education and whether they can be evaluated as a supportive tool in university exam preparation processes. The study sheds light on the possibility of using GAI models in the personalized education field that emerged with the rise of GAI models in the field of education. Further studies in this field may help develop new strategies for the integration of GAI models into the education system.

Keywords —generative AI, chatbots, chatGPT, gemini, copilot

Üretken Yapay Zeka Modellerinin (GPT-4o, Gemini, Copilot) YKS/TYT Sınavındaki Performansı: Karşılaştırmalı Bir Çalışma

***Özet*—** Bu çalışmada üretken yapay zekâ (GAI) modellerinin eğitim alanındaki potansiyeli değerlendirilmeye çalışılmıştır. Benzetimli bir sınav ortamı planlanmış ve GPT (Generative Pre-trained Transformer) ailesi modellerinin üniversite giriş sınavlarında sorulan sorulara verdiği yanıtlar baz alınarak hata analizi yapılmıştır. Çalışmada GPT, Gemini ve Copilot modelleri Türkiye'de Yükseköğretim Kurumları Sınavı (YKS)/ Temel Yeterlilik Sınavı (TYT) oturumunda yer alan sorularla test edilmiştir. Bu modellerin ürettiği yanıtlar hem doğruluk oranları hem de akademik derinlik ve tutarlılık açısından analiz edilmiştir. GPT-4o %50.8 doğruluk elde ederken, Gemini %43.3 ve Copilot %42.5 doğruluk elde etmiştir. Bulgular, GAI modellerinin eğitimde bir öğrenme yardımcısı olarak kullanılabilme potansiyeli ve üniversite sınavlarına hazırlık süreçlerinde destekleyici bir araç olarak değerlendirilip değerlendirilemeyeceği konusunda önemli ipuçları sunmaktadır. Çalışma, GAI modellerinin eğitim alanında yükselişiyle ortaya çıkan kişiselleştirilmiş eğitim alanında kullanılabilme ihtimaline ışık tutmaktadır. Bu alanda yapılacak daha fazla çalışma, GAI modellerinin eğitim sistemine entegrasyonu için yeni stratejiler geliştirmeye yardımcı olabilir.

***Anahtar Kelimeler*—** üretken yapay zeka, chatbotlar, chatGPT, gemini, copilot

1. INTRODUCTION

Large language models (LLMs) in AI are neural network models trained on large amounts of textual data that perform well in natural language processing tasks [1]. They enable models to understand, interpret, and respond to speech. Generative AI (GAI) is a subset of AI that focuses on creating new content by learning from existing data. It uses algorithms to generate text, images, music, and other forms of media based on natural language prompts submitted by users. The core technologies behind generative AI include Generative Adversarial Networks (GANs) and transformer models like GPT-4. Generative AI has a wide range of applications, including content creation, image generation, music composition, and more, making it a powerful tool for users.

According to August 2024 data from Liu and Wang's study "Who on Earth Is Using Generative AI?", ChatGPT-4o receives 82.5% of total web traffic among generative AI tools. It is followed by Gemini, Poe, Perplexity, Claude, DeepAI, and Copilot [2].

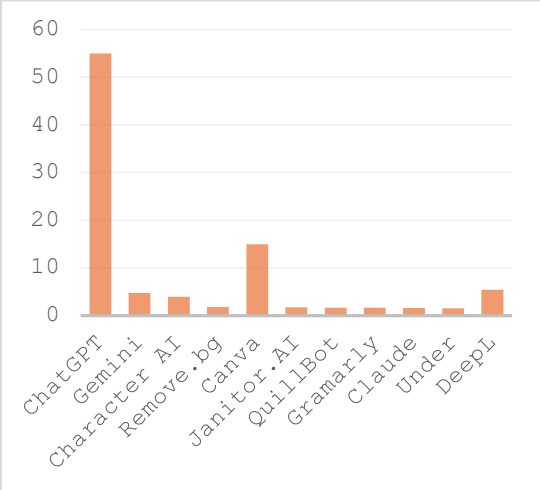


Figure 1. The most popular AI tools in 2024[3].

The analysis result in Figure 1 was obtained based on data collected from two reliable web analytics platforms, Semrush and Similarweb, as of August 2024. The most frequently used AI tool by market share was ChatGPT at 54.96%, followed by Canva at 14.92%, DeepL at 5.38%, and Gemini at 4.75% [3]. However, these values may change over time depending on usage patterns.

ChatGPT, developed by OpenAI in 2022, is a striking example of generative AI models. Generative AI is a complex artificial intelligence system that is capable of understanding, generating, and transforming human language through advanced machine learning techniques such as deep learning and neural networks [4]. These systems are trained on large amounts of data, continuously improving their performance in natural language processing tasks. Models such as ChatGPT have been successfully used in many areas such as natural language understanding, text generation, machine translation, and question-answer systems [5]. However, what these tools can do is not limited to just text generation. They can also generate digital content, including images, audio, and natural language, using GAI models [6].

With the increasing use of AI chatbots such as ChatGPT, these technologies have attracted significant attention in many fields, including higher education[7]. Numerous studies in this field have examined the potential of AI in education and have revealed the increasing interest in ChatGPT and similar technologies, especially in higher education [8-11]. In many studies, these AI tools have been used for reinforcing topics, preparing for exams [12-18], learning a different language [19-20], drawing new inferences using existing data [21-22], and personalized training [23]. There are also studies suggesting that ChatGPT can be used to prepare for exams [24-26]. These studies also state that these models can be used to support exams, but caution should be exercised. There are also studies conducted to generate sample questions for exam preparation [16, 27-28]. In general, ChatGPT 3.5, GPT4 Turbo or GPT-4o versions, Claude and Gemini models have been selected for comparison in studies. Often the procedures were repeated with these GAI tools and results were obtained.

Some studies have examined the accuracy of LLMs. Mavrych et al. analyzed the adequacy and accuracy of LLMs such as ChatGPT-4, ChatGPT-3.5-turbo, ChatGPT-3.5, Copilot, PaLM, Bard and Gemini in answering medical multiple choice questions and creating clinical scenarios and tests for upper extremity topics in Gross Anatomy course for medical students. ChatGPT-4 achieved the best result with 60.5% correct answer rate, followed by Copilot with 42.0% and ChatGPT-3.5 with 41.0%, followed by ChatGPT-3.5-turbo with 38.5%, Google PaLM 2 with 34.5% and Bard with 33.5%. It was concluded that despite technological advances, LLMs are not mature enough to fully assume the role of a teacher or facilitator in Gross Anatomy course; however, they can be valuable tools for medical educators [29].

Farhat et al.'s study evaluated the performance of 3 different LLMs, namely GPT-3.5, GPT-4 and Google Bard, in answering questions related to NEET-2023 exam. It was found that GPT-4 passed the entrance exam test with 42.9%, GPT-3.5 met the proficiency criteria with 20.7%, but Bard failed the

proficiency criteria with 16.4% and failed the test. The study highlights the importance of LLMs in terms of important exams and their positive impact on education, and encourages the responsible and conscious use of these models in various learning environments [16].

Abu-Haifa et al.'s study analyzed the performance of Copilot, ChatGPT, and GPT-4 on Graduate Record Examination (GRE) questions. LLMs were assessed with 137 questions on different quantitative reasoning styles and 157 questions on verbal categories, and their proficiency in complex language comprehension and visual questions was examined. The results show that ChatGPT is more prone to arithmetic errors, while Copilot and GPT-4 are more prone to conceptual errors. GPT-4 performed best on complex and visual-based tasks. The study highlights the potential benefits of these tools in GRE exam preparation [30].

Koç and Öztiryaki evaluated advanced large language models (LLMs) such as Gemini Advanced, GPT-4, Copilot, and GPT-3.5 to analyze the accuracy of answers to questions related to medical imaging systems. A total of 400 questions were divided into four categories: X-ray, ultrasound, magnetic resonance imaging, and nuclear medicine, and the models were analyzed according to the percentage of correct answers with a zero-command approach. The accuracy rates were calculated as 86.25% (Gemini Advanced), 84.25% (GPT-4), 77.5% (Copilot), and 59.75% (GPT-3.5), respectively [31].

Öter and his colleagues evaluated the ÖSYM/KPSS exam questions of the last three years with ChatGPT. They achieved a success rate of 51% for the KPSS exam in 2021, 50% in 2022 and 40% in 2023. The study discussed how publishing all ÖSYM questions in 2021 and only 10% of the questions in the last two years affected the results and the effect of the prompt for questions asked to ChatGPT[32].

We investigate the exam performance of three popular GAI models, GPT-4o, Gemini and Copilot, in the TYT session of YKS. Our aim is to evaluate these models for university entrance exams and to investigate their potential to be used as a supplementary learning tool for the TYT exams to be held in the coming years.

1.1. ChatGPT

ChatGPT was released by OpenAI in November 2022 [33] and gained huge popularity with over 100 million users by January 2023 [34]. The model, based on the GPT-3.5 architecture, was made available to all users for free. In March 2023, a paid version, ChatGPT Plus (GPT-4), was released, offering faster response times and additional features for a monthly subscription fee of $20 [35]. In May 2024, the GPT-4o version, equipped with human-like features, was introduced [36].

While GPT-3.5 works with 175 billion parameters, GPT-4 is reported to have 1.7 trillion parameters [37]. The parameters are used to adjust the tone, style, and creativity of the AI's responses. Furthermore, GPT-3.5 can handle up to 3,000 words, while GPT-4 and GPT-4 Turbo can handle up to 25,000 and 96,000 words respectively [38]. ChatGPT records user responses and history as a conversation thread, but in GPT-4 Turbo, sharing this data is left to the user's discretion. GPT-4o can handle multiple inputs, including text, audio, images, and video, offering near-human response times and significant improvements, especially in visual and auditory understanding [39]. The model, which also performs well in languages other than English, is available free of charge to all users for a limited number of queries. As of August 2024, ChatGPT has 100 million weekly and 180.5 million monthly active users. Daily traffic has exceeded 100 million since the announcement of GPT-4o [40].

ChatGPT's data included a data set up to September 2021, and there was a problem accessing the latest data due to lack of internet connection[41]. The web search application, which was offered to premium subscribers as of September 2024, was opened to all users by the end of the year.

1.2. Copilot

In February 2023, Microsoft introduced a new version of its search engine Bing to compete with ChatGPT. At launch, the AI chatbot was called Bing Chat. However, during Microsoft Ignite in November 2023, Microsoft decided to rebrand the chatbot and name it Copilot instead[42].

You can ask Copilot questions and get detailed, human-like answers with footnotes and links back to sources. Since Copilot is connected to the internet, it can provide up-to-date information as it can access data via the web. It can help you with creative tasks such as writing poems, essays, or songs. It can solve complex math or coding tasks, and even create images from text using Microsoft Designer's Image Generator. It does this using Microsoft's AI image generator, DALL-E, OpenAI. Moreover, the most important feature of Copilot is its ability to generate answers using GPT-4 Turbo [43].

Like ChatGPT, Microsoft Copilot lets you create text, essays, letters, summarize content, write code, and answer complex questions. It also has internet

access, allowing it to provide up-to-date answers on current events. You have three conversational style options for how Copilot responds to your requests: creative, balanced, and precise. Each of these formats is self-describing: “more creative” will give you original and creative answers. “more balanced” is similar to ChatGPT, which is an informative and friendly chat with a human-like response. If you choose a balanced approach, the model will produce a more neutral output by considering both aspects. However, if you want to get more specific and precise information, it is recommended to make the model’s style more precise [44]. The Copilot app, developed by Microsoft, also comes in two versions: a free version and a Copilot Pro version, which comes with Office 365+ Microsoft applications for $20 per month [43].

1.3. Gemini

Gemini, previously known as Bard, is an advanced AI chatbot developed by Google, released on March 21, 2023, and based on LaMDA technology [43]. Gemini 1.0 was trained on Google's AI-optimized infrastructure using TPU v4 and v5e, while the next-generation TPU v5p accelerates its development and improves training for large-scale generative models [44].

Gemini 1.5 Flash is currently available in 230 regions in over 40 languages and features 32,000 fragmented context windows to increase responsiveness and efficiency[45]. Both Gemini 1.5 Pro and Flash offer context windows of up to one million tokens, providing nearly perfect recall for tasks involving long documents, extensive code bases, or multimedia data [35].

Designed for natural language processing, Gemini provides more comprehensive and up-to-date answers through Google’s knowledge graph. The AI Premium plan, which includes Google services, is priced at $20 per month, with a free version available [36].

2. MATERIAL AND METHODS

Entrance to higher education institutions in Türkiye is based on a three-session exam system called YKS. The first of these sessions, TYT, is an exam that aims to measure the basic academic skills of candidates and includes questions from Turkish, Mathematics, Social Sciences and Science. TYT evaluates the general academic competence of candidates with 120 questions in 135 minutes and constitutes 40% of the total score of YKS. Candidates have the right to choose two-year associate degree programs by taking only the TYT exam [47].

2.1. Models

In the study,

- GPT-4o v1.0 (Release date: June 2024; available with OpenAI Plus subscription),
- Gemini 1.5 Flash (Release date: April 2024; free access),
- Microsoft 365 Copilot Chat v2.0 (Release date: May 2024; free access via Microsoft Edge).

Reasons for choosing both paid and free versions; to understand whether there is a significant statistical difference between these models and there is a budget constraint due to using the paid version.

The models have not been trained with similar questions from user accounts before. However, this does not guarantee that the models have not encountered similar questions before. The study was designed considering that the models have not been trained with similar questions. The questions were asked in exactly the same way as given in the ÖSYM guide and were used only to get the correct answer.

Since the questions are routinely released to the public every year, it is considered a public dataset and no institutional permission is required. All questions were asked to the models on 07/19/2024.

2.2. Tools and Materials

Correct answer rate, error types (comprehension error, calculation error, lack of information) and solution time were used as evaluation criteria.

If the models did not arise from a lack of information but from misinterpreting the question, these questions were marked as “comprehension errors”, if there were operational errors due to a lack of information, these errors were marked as “lack of information errors”, and errors resulting from mathematical operations were marked as “calculation errors”.

2.3. Procedure

A file was created containing each of the YKS/TYT questions as images. In addition to the questions, the correct answers were also kept in a separate file. A new file was also created containing the text of each question in case there was a problem with the image files. Each question was given as input to the models. As output, we prompted the models to produce texts containing the correct answers to the questions. We obtained a statement that includes both the correct answer option and information about the question.

2.4. Response Evaluation:

In multiple choice questions, the answers given by the models were evaluated by comparing them with the correct answer key. Incorrect answers were classified according to error types. Error types included comprehension errors (misunderstanding the question), calculation errors (performing mathematical operations incorrectly), errors of lack of knowledge (not having the necessary information) and other errors.

2.5. Data Analysis

The correct answer rate was calculated for each model. Accuracy is defined as the percentage of correct answers obtained by a model. In this research, accuracy was obtained using the following formula:

Accuracy = Correct Answers / Total Answers

The chi-square test was applied to understand whether there is a significant difference between the accuracy rates of the models. The chi-square test evaluates whether the difference between the observed frequencies and the expected frequencies is statistically significant.

It is calculated with the formula: $X^2 = \sum \frac{(F_i - E_i)^2}{E_i}$

Here F_i represents the actual frequencies and E_i represents the expected frequencies.

Hypotheses in our study:

H_0=There is no relationship between ChatGPT, Gemini and Copilot, the variables are independent.

H_1=There is statistically significant relationship between ChatGPT, Gemini and Copilot.

While the chi-square test only answers the question "is there a difference?", Cohen's f calculation method used to measure the magnitude of the difference.

$f = \sqrt{\frac{n^2}{1-n^2}}$ represents the ratio of the variance between the groups to the total variance: $n^2 = \frac{SS_{between}}{SS_{total}}$

In the study, the frequency and distribution of different error types were examined and the average time spent by the models to answer each question was calculated with the Python program written. For chi-square test analysis, the website socscistatistics used.

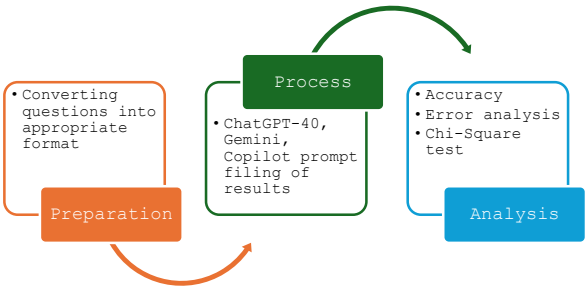


Figure 2. Basic Workflow Followed in the Study

Figure 2 shows the stages of the study. First, the test questions were converted into image files for each question during the preparation process and recorded in a word file. Before the questions were asked, the models were informed that they would be given test questions from a certain area and were asked to indicate only the correct answer. The command given to the models in Turkish was as in Figure 3:"I will give you the test questions in order, I want you to give the correct answer. Are you ready?". The questions were given to the models one by one and the answers were recorded for evaluation. The answers produced by the model were accompanied by explanations on how the problem was solved. Figure 4 shows a sample query. The answer produced by the model included information about the question and how the problem was solved. In cases where no answer was received, the answer given by the model was marked as incorrect. In questions that included visuals, unlike the text, GPT-4o, Copilot and Gemini were able to perform visual processing, but they could not always reach the correct result for the question. The answers obtained were recorded in the excel file as question number, GAI model, correct/wrong answer, response time and error type. In the analysis section, the accuracy of each model, error analysis, general and area-based Chi-square tests and Cohen's f effect factor calculation were performed.

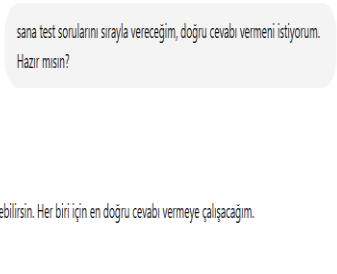


Figure 3. First Prompt for models

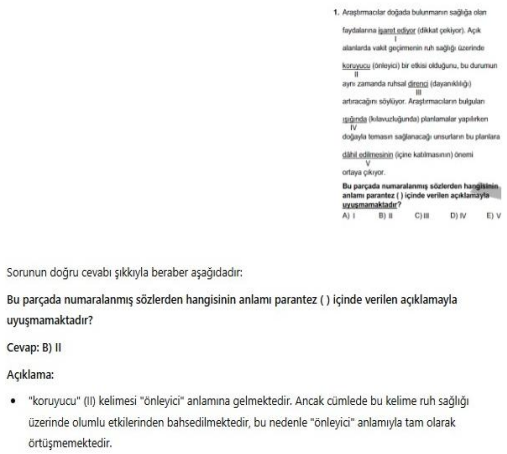


Figure 4. An example prompt for ChatGPT

3. RESULTS AND DISCUSSION

2024-YKS/TYT was conducted by ÖSYM on June 8, 2024, and the basic question booklets and answer keys were published the next day [46]. The success levels of ChatGPT, Gemini and Copilot generative artificial intelligence tools in the TYT exam conducted by ÖSYM are shared in Table 1. The times were calculated by writing a Python program and saved in a file. The average of the answering times of all questions for each model was rounded. When we look at the times, we see that they work in very similar times. In this study, only the first answers produced by GPT-4o, Gemini and Copilot were evaluated without using the reconstruction function.

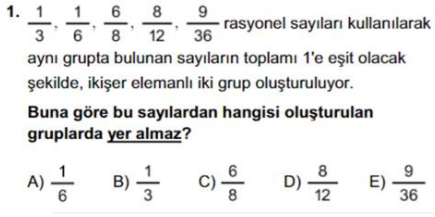
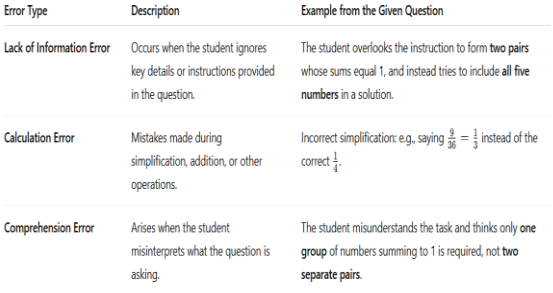


Figure 5. An example math question for ChatGPT

How a math question given in Figure 5 can be interpreted with different errors is shown in Figure 6. With this question example, it is also clarified how the errors encountered in different models are grouped



Error Type	Description	Example from the Given Question
Lack of Information Error	Occurs when the student ignores key details or instructions provided in the question.	The student overlooks the instruction to form **two pairs** whose sums equal 1, and instead tries to include **all five numbers** in a solution.
Calculation Error	Mistakes made during simplification, addition, or other operations.	Incorrect simplification: e.g., saying $\frac{9}{36} = \frac{1}{3}$ instead of the correct $\frac{1}{4}$.
Comprehension Error	Arises when the student misinterprets what the question is asking.	The student misunderstands the task and thinks only **one group** of numbers summing to 1 is required, not **two separate pairs**.

Figure 6. An example error types for given question with ChatGPT

Table 1. Types of accuracy and error of artificial intelligence tools in the university entrance exam

Models	Correct Answer Rate (%)	Average Solution Time (sec)	Types of Errors
GPT-4o	50.8	2.5	Comprehension errors, calculation errors
Gemini	43.3	3.0	Lack of information errors, Comprehension errors, calculation errors
Copilot	42.5	2.0	Lack of information errors, calculation errors, Comprehension errors,

The overall accuracy of the models is as follows: GPT-4o achieved a maximum accuracy rate of 50.8% by correctly identifying 61 out of 120 answers. The reason for these low results is that the data was last updated in September 2021, and the questions were encountered for the first time. It is clear that these values will be higher if this study is repeated in the near future.

Gemini answered 52 correct answers with 43.3% accuracy and Copilot answered 51 out of 120 answers with 42.5% accuracy. The difference in accuracy with GPT may also be due to the more advanced features of the paid versions. In order to understand whether there is a statistically significant difference between ChatGPT, Gemini and Copilot, a Chi-square test was performed and the p-value was found to be 0.360897. In this case, the null hypothesis (H_0) is not rejected, meaning that there is no statistically significant relationship between ChatGPT, Gemini and Copilot. In other words, these three tools were evaluated independently of each other.

Subject-based accuracy rates of the models are given in Table 2.

Table 2. Subject-based accuracy rates and numbers

Lesson	Number of questions	GPT-4o		Gemini		Copilot	
		%	Number of Correct	%	Number of Correct	%	Number of Correct
Turkish	40	50	20	**52.5**	21	47.5	19
Maths	40	**35**	14	15	6	25	10
History	5	**80**	4	**80**	4	40	2
Geography	5	**80**	4	20	1	20	1
Philosophy	5	40	2	**80**	4	60	3
Religion	5	**60**	3	**60**	3	**60**	3
Physical	7	**57**	4	10.4	1	20.8	2
Chemical	7	57	4	**100**	7	85.7	6
Biology	6	**75**	6	62.5	5	62.5	5
Total	120		61		52		51

TYT exam consists of 40 Turkish, 40 Mathematics, 5 each from History, Geography, Philosophy and Religion, 7 Physics, 7 Chemistry and 6 Biology. Table 2 shows these question numbers, the accuracy rates of the answers given by these tools to these questions, and the number of correct answers. The highest accuracy in Turkish is found in Gemini with 52.5%, in Mathematics with GPT-4o with 35%, in History with Gemini and GPT-4o with 80%, in Geography with GPT-4o with 80%, and in Philosophy with Gemini with 80%. In Religion, GPT-4o, Gemini and Copilot are tied with 60%. The highest accuracy in Physics was obtained with 57% in GPT-4o, in Chemistry with Gemini with 100%, and in Biology with 75% in GPT-4o.

Thirty-two of the TYT exam questions contained visual content. GPT-4o was able to answer 10 of these questions correctly, while Gemini answered only 4 and Copilot answered 5. While these generative models are good at deciphering characters in images , they struggle to understand and solve the question. Gemini reported many math questions as incorrect. The watermarks in the background of the ÖSYM question guide were particularly challenging for the Gemini tool, as they were superimposed on the images.

GPT-4o showed an average performance in the Turkish field with 20 correct and 20 incorrect answers to the questions. Gemini answered 21 questions correctly, while Copilot answered 19 questions correctly. The incorrectly answered Turkish questions were related to sentences that distorted the meaning and paragraph analysis. According to the Chi-square test, the p value was calculated as 0.904837. In this case, since the p value is greater than 0.05, the null hypothesis (H_0) *cannot be* rejected, meaning that there is no statistically significant relationship between the variables.

In math, GPT-4o answered 14 questions correctly and 26 questions incorrectly. Gemini answered only 6 questions correctly. For many questions, it responded “The question cannot be solved with the given information” and failed to solve them. Copilot answered 10 questions correctly. There are a total of 22 questions in this field that contain images, and most of the questions that GPT-4o, Gemini, and Copilot answered incorrectly were questions that contained images. This result suggests that GPT-4o and other tools need to improve their knowledge-based interpretation capabilities in mathematical questions that contain visual content. According to the Chi-square test in Math, the p value was calculated as 0.118442 ($p \leq 0.05$). In this case, the null hypothesis (H_0) *cannot be rejected*, meaning that there is no statistically significant relationship between ChatGPT, Gemini and Copilot.

In mathematics, especially geometry problems, GPT-4o sometimes produced results that were not among the given answer choices. In this case, the question answer was interpreted as an incorrect solution and a lack of information. When the model was told that the answer was wrong for testing purposes, it solved the question again and still produced an answer that did not match the choices. When the question was asked again, it solved the question successfully on the third try. For example, on a geometry question, GPT-4o initially produced a result that was not listed in the choices. After highlighting this, it produced an incorrect answer, in which case the answer was recorded as incorrect. However, when the question was asked again with the correct answer, it solved the question as expected. Then, when the same question was asked again after some time, GPT-4o solved it correctly. This behavior suggests that the model was either not trained for this particular scenario beforehand or that there was no relevant data in the dataset .

GPT-4o answered 13 out of 20 questions in Science correctly. Some of the questions it answered incorrectly were in Physics and some in Chemistry. It made almost no mistakes in Biology. This shows that GPT-4o is more successful in visual content questions in Science than in Math. Gemini answered 12 questions correctly in Science, while Copilot answered 13 questions. It is also interesting that Gemini answered all of the Chemistry questions correctly. This suggests that there may be a data set about chemistry beforehand. GPT-4o answered 4 questions correctly in Chemistry, while Copilot answered 6 questions. While Gemini had difficulty solving the questions in Physics, Copilot didn't have any difficulty. While there were 7 visual questions in this section, GPT-4o could correctly solve 4 questions, Gemini 1 and Copilot 2 of them. According to the Chi-square test in Science, the p value was calculated as 0.930745 ($p > 0.05$). In this case, the null hypothesis (H_0) *cannot be* rejected, meaning that there is no statistically significant relationship between ChatGPT, Gemini and Copilot.

GPT-4o answered 13 out of 20 questions in the Social Sciences section correctly. Gemini and Copilot answered 12 and 9 questions correctly, respectively. The questions in this area consisted of 5 questions in History, 5 in Geography, 5 in Philosophy, and 5 in Religion. In History, GPT-4o and Gemini answered 4 questions correctly, and Copilot answered 2 questions. In Geography, GPT-4o answered 4 questions, and Gemini and Copilot answered 1 question correctly. In Philosophy, GPT-4o answered 2 questions, Gemini answered 4 questions, and Copilot answered 3 questions correctly. Finally, in Religion, they answered the same number of questions correctly, 3 questions. GPT-4o and Copilot answered the same questions correctly, while Gemini answered different questions correctly. There were 3 visual questions in Geography. While GPT-4o could answer 2 of these questions correctly, the others could only answer 1 correctly. There were comprehension errors in the incorrectly answered questions because they required interpreting the given text rather than measuring knowledge. According to the Chi-square test for the social sciences, the p value was calculated as 0.802695 ($p \leq 0.05$). In this case, the null hypothesis (H_0) *cannot be rejected*, meaning that there is no statistically significant relationship between ChatGPT, Gemini and Copilot.

When the answers to the questions were examined, GPT-4o produced a better result than Gemini and Copilot. Despite their similar performance in the Turkish test, GPT-4o was almost twice as successful as Gemini and Copilot, especially in mathematics. According to the chi-square test conducted to see if there was a statistical basis for this, the p value was found to be 0.904837 ($p \leq 0.05$). In this case, the null hypothesis (H_0) *cannot be rejected*, meaning that there is no statistically significant relationship between the variables.

In addition, GPT-4o performed twice as well as Gemini and Copilot in visual content questions. When we look at the subject-based social field, we see that GPT-4o is more successful in History and Geography, Gemini in Philosophy; In the field of science, GPT-4o is more successful in Physics and Biology, and Gemini in Chemistry.

Table 3 shows the statistical analysis of model performances in the study.

Table 3. Statistical Analysis of Model Performances

Sections	p-value	Interpretation	Cohen's f	η^2	Comment
All	0.360897	no difference	0.08	0.006	Weak effect
Turkish	0.904837	no difference	0.08	0.005	insignificant difference
Social Sciences	0.802695	no difference	0.22	0.046	Small effect
Maths	0.118442	no difference	**0.65**	**0.297**	Big effect
Sciences	0.930745	no difference	0.05	0.003	insignificant difference

As a result of the chi-square analysis, no significant difference was found between the GPT-4o, Gemini and Copilot models ($p > 0.05$). This shows that the models have similar accuracy rates and the null hypothesis (H_0) was accepted because $p>0.05$ was found in all sections. In particular, the p-values obtained in the general assessment ($p = 0.361$) and subtests (Turkish: $p = 0.905$; Social Sciences: $p = 0.803$; Science: $p = 0.931$) support the hypothesis (H_0) that the models work as independent systems. This result indicates that there is no statistically significant difference between the models. However, considering that statistical significance may be affected by the limited sample size, Cohen's f effect size was calculated to measure the magnitude of the difference. This metric provides a standardized measure of the difference between groups in the analysis of variance (ANOVA), and is interpreted as small between 0.10-0.25, medium between 0.25-0.40 and large>0.40 [47]. "The large effect size observed in the Math section ($f=0.65$, $\eta^2=0.297$) confirms the significant performance superiority of

GPT-4o over the other models. This result suggests that the computational capacity of the paid model (GPT-4o) provides a critical advantage in mathematical problem solving [37]. However, the fact that p=0.118 indicates that this difference does not exceed the threshold of statistical significance ($p<0.05$). This discrepancy suggests the possibility of a Type II error due to the limited sample size (n=40).

The small effect in the social sciences section (f=0.22) indicates that the 20-point absolute difference between GPT-4o and Copilot is limited by methodological factors (watermarked images). The negligible effect sizes in all other sections ($f\leq0.08$) support the interpretation that the performances of the models are generally equivalent.

Especially in the subtests consisting of 5 questions (history/geography), the high η^2 values should be interpreted with caution due to limited statistical power. The findings suggest that differences in exam performance between GAI models, although not statistically significant, may be important in a pedagogical context.

Table 4 shows the number of questions in the YKS session based on the field, the average correct number of GAI models and candidates. The average number of questions for candidates in 2024 was calculated as the average of all students who took the exam [48]. When we examine the table; Gemini performed better than the candidates in the Turkish field test, while all GAI models were much more successful than the candidates in the Social, Math and Science field tests. It is an important inference that ChatGPT-4o was 1.5 times more successful than the candidates in Social, 2 times more successful in Math and 4 times more successful than the candidates in Science. Gemini and Copilot, which are free, also achieved good results. Copilot was below average in the Turkish test, while Gemini was below average in the Science test. However, these figures are quite close to the average. These results reveal the potential of these models to be used as an aid in the preparation phase for the exams.

Table 4. Area-based correct numbers

Exam Name	**Number of Questions**	Candidates average question	**GPT4o**	**Gemini**	**Copilot**
2024/TYT-TURKISH	**40**	20.17	20	21	19
2024/TYT-SOCIAL	**20**	8.39	13	12	9
2024/TYT-MATHEMATICS	**40**	6.92	14	6	10
2024/TYT-FEN	**20**	2.86	13	12	13

According to the accuracy analysis obtained; the possible scores that could have been obtained if GPT-4o, Gemini and Copilot had taken the TYT exam were obtained by a scoring robot providing service over the internet, as in Table 4. The net numbers were calculated by subtracting one-fourth of the number of wrong answers from the number of correct answers.

According to Table 5, GPT-4o received 278.96810 points with a net of 45, while Gemini received 253.05641 points with a net of 36.25, and Copilot received 246.02143 points with a net of 33.75. In the TYT exam, which was taken by 2755277 people, GPT-4o entered the 21.68% percentile, Gemini 30.20%, and Copilot 37.34%. When these figures are examined, it is seen that GPT-4o is clearly ahead.

Table 5. GPT-4o, Gemini and Copilot TYT scores and results

	Turkish	**Social**	**Mathematical**	**Science**	**TYT Score (excluding school score)**	**TYT ranking (2,755,277 out of 2)**	**Percentage**
GPT-4o	15	11.25	7.5	11.25	278.96810	597,416	21.68%
Gemini	16.25	10	0	10	253.05641	924,304	30.2%
Copilot	13.75	6.25	2.5	11.25	246.02143	1,028,780	37.34%

The success scores received from the school have not yet been added to these scores calculated in the table. A score between 30-60 will be added to these raw scores depending on the success in school. Students can then be placed in the two-year associate degree programs they prefer with the score obtained.

A non-academic study was conducted to measure the performance of ChatGPT-4o with YKS exam questions in 2024, but it was not stated how the analysis was done. GPT-4o scored 28.75 net in Turkish, 16.25 net in social sciences, 2.5 in mathematics and 5 in science test in TYT. In the Field Proficiency Exams (AYT), it scored 35 net in social sciences, 2.5 net in science and -5 net in mathematics. Accordingly, GPT-4o scored 469 in verbal and 239 in numerical, for an equal weight of 347. With these scores, it was estimated that GPT-4o could be placed in Boğaziçi University Turkish Language and Literature Department and Middle East Technical University (METU) History Department[48].

When correct question values in our study are examined; GPT-4o was able to get 15 in Turkish, 11.25 in social studies, 7.25 in mathematics and 11.25 in science in the TYT session. When the results of the two studies are compared, it is seen that there are differences. While there was a significant decrease in Turkish and social studies, there was an increase in the net number in mathematics and science. When the possible reasons for this are investigated; similar questions were asked before, the questions were given with the copy-paste method instead of the image processing method, the watermarks in the background of the booklet were deleted, the fine-tuning was done with the GPT 4.0 application and the date the questions were asked. In the other study [49], the AYT exam was also evaluated as an extra. In our study, only the TYT exam was evaluated with 3 different GAI.

A website used the GPT 4 language model to solve the 2021 YKS/TYT exam. It was seen that the model made 27.75 net in the Turkish , 18.75 net in social sciences questions, 13.75 net in science section and 17.50 net in Mathematics questions [50]. Accordingly , GPT-4 received 419 points in the 2021 TYT exam. When ÖSYM's data was examined; It showed that the model was 99.50% better than the participants who took the exam in Turkey that year. In this example given, there is no explanation about the conditions and how this process was done. It is seen that the number of correct answers is high. However, in questions containing visuals, the given answer is counted as correct. This does not reveal how accurately the results were analyzed. As a result, it turns out that the ChatGPT model is more successful against humans. In this study, solving the questions from previous years with ChatGPT in order to prepare for the exam environment can be considered as an important development for the personalized education model. However, it is necessary to evaluate the hallucination status of LLMs due to lack of information, especially for questions answered incorrectly .

As another output of our study, let's take a look at how these models can play a role in people's learning processes. During the preparation process for university exams, AI-supported tools can provide students with personalized educational support and provide an efficient learning process. These tools offer features that can be customized according to students' various needs. Below are items on how these tools can be used:

- Subject Explanation: Students can get explanations on topics they have difficulty with and learn complex topics in a simpler way.
- Practice Questions and Tests: Students can practice taking tests by asking exam-style practice questions. Tools can solve questions instantly and provide feedback.
- Feedback and Evaluation: It can determine the correct and incorrect answers given by the students and provide development by explaining the missing or incorrect points.
- Study Plans: Students can create customized study plans that are appropriate to their current level. These plans can provide guidance on which topics need further study.
- Time Management: During the exam preparation process, they can get suggestions on how much time should be spent on which topics, and obtain strategic advice on how to use time most efficiently for the exam.
- Motivation and Support: Students can find morale and motivation during the preparation process and receive suggestions for managing exam anxiety.

These tools provide students with the opportunity to effectively prepare for exams, correct their deficiencies, and improve their time management skills.

However, there are also some disadvantages to using such models. First of all, students may become overly dependent on these tools and have difficulty developing cognitive skills such as critical thinking and problem solving. In addition, it should be noted that GAI models can sometimes provide incorrect or incomplete information, which can lead to misleading results in the learning process. Another important disadvantage is that such systems do not provide equal access to every student; some students may not be able to benefit sufficiently from these tools due to economic or technological limitations.

Finally, AI-supported learning processes may have some pedagogical deficiencies as they can not completely replace teacher-student interaction.

4. LIMITATIONS

One of the most important limitations of this study is the use of different versions of GAI models. While the paid version of GPT-4o was used in the study, the free versions of Gemini and Copilot were preferred. This prevents a fair comparison between the models, as the paid version of GPT-4o provides more advanced processing power, faster response times, and access to more comprehensive data sets compared to the free versions. This difference may be one of the reasons why GPT-4o achieved an accuracy rate of 50.8%, while Gemini remained at 43.3% and Copilot at 42.5%.

Another important limitations of this study is that since the training data of the GAI models used is not publicly available, the possibility that the models have seen similar exam questions before can not be completely excluded. However, the possibility of directly including content such as the 2023 TYT exam questions, which are only published in PDF format, copyrighted, and watermarked by ÖSYM, is quite low. Nevertheless, the fact that this possibility can not be definitively verified constitutes an important methodological and ethical limitation.

The research was conducted in July 2024, and the results of the study may change since the models were updated after this date. Due to the fact that the 2024 YKS/TYT questions are public and the GAI models are constantly updated, there is a possibility that these questions will be included in the training set of the models in later periods. This may lead to different results in similar studies in the future and may negatively affect the reproducibility of the study.

The large effect size in mathematics (f=0.65) may not have reached the significance threshold (Type II error) due to the sample size (n=40) limiting the statistical power. A minimum of 120 questions is recommended for the mathematics section in future studies.

Although Cohen's f calculations are not affected by sample size, the confidence intervals of the effect size widen, especially in subtests consisting of 5 questions (history/geography). Therefore, high f values in small subgroups (e.g. f = 0.51 in philosophy) should be interpreted with caution.

Technically, the important limitations of the study are the ÖSYM watermarks that make it difficult to analyze visual questions, the deficiencies in the graphic and shape interpretation capabilities of the GAI models, and the low success rates in questions with visual content. In addition, the requirement of Copilot to start a new session every 30 commands, the hourly usage quotas in the free versions, and the occasional usage stop messages of the models have negatively affected the data collection process.

Methodologically, evaluating only the first answers for each model and not using the reconstruction function ignores the possibility of obtaining different results when the same question is asked again. The fact that GAI models sometimes produce realistic but false information (hallucinations) and that answers such as "This question cannot be solved with the given information" indicate deficiencies in the training set reveal the risks of relying solely on GAIs in critical decision-making processes. In terms of generalizability, evaluating only the YKS/TYT exam creates limitations in terms of generalizability to other standard tests. Conducting the study only in the context of the Turkish education system and conducting the evaluation at a single point in time provides limited representative power in a dynamic environment where models are constantly evolving.

From an ethical and pedagogical perspective, the potential for over-reliance on GAI models to negatively impact critical thinking skills, the fact that not all students have equal access to these tools due to economic and technological constraints, and the fact that AI-supported learning processes can not fully replace traditional pedagogical interactions are important limitations. To overcome these limitations, it is recommended that future studies use equivalent (paid/free) versions of all models, include different GAI models (Claude, Llama, Ernie), conduct retests at multiple time points, and include open-ended questions in the assessment.

5. CONCLUSION AND SUGGESTIONS

This study presents critical findings regarding the integration of generative artificial intelligence (GAI) in education by systematically evaluating the performance of GPT-4o, Gemini and Copilot models in the YKS/TYT exam. The overall accuracy rates were GPT-4o (50.8%), Gemini (43.3%) and Copilot (42.5%), while chi-square tests showed that there was no statistically significant difference between the models ($p>0.05$). This indicates that each model functions as independent systems [30,31]. Especially in the field of mathematics, the fact that GPT-4o (35%) performed 2-3 times more than the other models revealed the determining role of parameter size and computational capacity [37,38]. On the other hand, low success rates were observed in questions with visual content (26.6% of TYT questions); visual noise caused by ÖSYM watermarks has revealed the limitations of the

models in their ability to interpret shapes and graphs [21,31].

The large effect size in mathematics (f=0.65) highlights the discipline-specific impact of model architecture in educational technology selection. In the preparation process for the university exam:

- High computational capacity models for mathematics (GPT-4o)
- Models with up-to-date data for knowledge-based fields such as chemistry (Gemini)
- For visually-based questions, models with strong multimodal integration should be preferred.
- However, the 17-23% hallucination rate observed in all models (Table 1) suggests that the outputs should not be used without teacher supervision.

Error analysis highlights significant risks to the academic use of the models. Gemini's response of "This question can not be solved with the given information" in 40% of the math questions is attributed to deficiencies in the training dataset [29,52], while GPT-4o's non-choice answers to geometry questions confirm the tendency of LLMs to produce realistic but fictional outputs (hallucinations) [51,54]. These findings point to the potential dangers of using GAIs alone in areas that require high accuracy, such as medical diagnosis or exam preparation [54]. Compared to human candidates, the models outperformed the student average in all disciplines (2.3x in social sciences, 4.5x in science, 2x in mathematics), and the science net of GPT-4o (11.25) was calculated as four times the student average (2.86) (Table 4). While these results support the potential of GAIs as personalized learning supporters [16,23], they should be interpreted with caution due to conflicting findings in the literature (e.g., studies suggesting that GPT-4o is in the 99th percentile [49,50] and methodological differences (visual loading techniques, first-response evaluation).

The following suggestions can be made for GAI integration in education:

Pedagogical Adjustments: Step-by-step solution explanations of the models can be used as complementary tools in the exam preparation process [12,27]. However, the risk of hallucination necessitates filtering the outputs with expert supervision [53]. To preserve students critical thinking skills, "teacher-AI collaboration" models should be designed [11,23] and school-based access centers should be established to reduce economic inequality [8].

Technological Improvements: Watermark removal algorithms and multimodal training (text + visual integration) should be developed for solving visual questions [21]. Parameter optimization and new generation processors such as TPU v5p should be used to improve the performance of free model versions (Gemini, Copilot) [44].

Future Research: Studies should be repeated with equivalent versions (paid/free) and different GAIs (Claude, Llama) [30]. The effect of dynamic updates on performance (after questions are added to the training set) should be monitored [29], and open-ended questions and long text analysis should be included in the evaluation [14].

In conclusion, while GAI systems have the potential to support individual learning processes, integrating these tools directly into exam preparation without addressing their limitations, such as the risk of hallucination, response variability, and confidence uncertainty, through human supervision and technological improvements, is pedagogically and ethically risky; striking this balance is critical for the sustainable integration of AI in education.

REFERENCES

[1] T., Eloundou, S., Manning, P., Mishkin, & D. Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models", arXiv preprint arXiv:2303.10130. 2023.

[2] Y. Liu, H. Wang. "Who on Earth Is Using Generative AI?", https://elements.visualcapitalist.com/wp-content/uploads/2024/09/1726222967151.pdf, 2024.

[3] Internet: D. Gewirtz/ZDNET https://www.zdnet.com/article/the-most-popular-ai-tools-of-2024-and-what-that-even-means/. 23.12.2024

[4] S. Bulut, "Üretken Yapay Zeka Teknolojisi: GPT-4o", 2024. *International Journal of Advanced Natural Sciences And Engineering Researches* (8)4: 380 – 387.

[5] A. Bozkurt, ChatGPT, "Üretken yapay zeka ve algoritmik paradigma değişikliği", Alanyazın, 4(1), 63-72. 2023

[6] Y., Cao, S., Li, Y., Liu, Z., Yan, Y., Dai, P. S., Yu, & L, "Sun, A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt", arXiv preprint arXiv:2303.04226. 2023.

[7] B. D., Lund, & T. Wang, "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?", Library hi tech news, 40(3), 26-29. 2023.

[8] D. R., Cotton, P. A., Cotton, & J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT", Innovations in education and teaching international, 61(2), 228-239. 2024.

[9] M. Firat, "How ChatGPT can transform autodidactic experiences and open education?", 2023.

[10] A. Iskender, "Holy or unholy? Interview with OpenAI's ChatGPT", *European Journal of Tourism Research*, 34, 3414-3414. 2023.

[11] P., Zhang, & G. Tur, "A systematic review of ChatGPT use in K-12 education". *European Journal of Education*, 59(2), e12599. 2024.

[12] M., Liu, T., Okuhara, Z., Dai, W., Huang, H., Okada, F., Emi, & T. Kiuchi, "Performance of Advanced Large Language Models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese Medical Licensing Examination: A Comparative Study", medRxiv, 2024-07. 2024.

[13] L. Lian, Comparative Study of GPT-4.0, "ERNIE Bot 4.0, and GPT-4o in the 2023 Chinese Medical Licensing Examination", 2024.

[14] J., Savelka, A., Agarwal, C., Bogart, & M. Sakr, From GPT-3 to GPT-4: On the Evolving Efficacy of LLMs to Answer Multiple-choice Questions for Programming Classes in Higher Education, **In International Conference on Computer Supported Education** (pp. 160-182). Cham: Springer Nature Switzerland, 2023.

[15] D. M., Katz, M. J., Bommarito, S., Gao, & P. Arredondo, "Gpt-4 passes the bar exam", Philosophical Transactions of the Royal Society A, 382(2270), 20230254, 2024.

[16] F., Farhat, B. M., Chaudhry, M., Nadeem, S. S., Sohail, & D. Ø. Madsen, "Evaluating large language models for the National Premedical Exam in India: comparative analysis of GPT-3.5, GPT-4, and Bard", JMIR Medical Education, 10, e51523, 2024.

[17] M. L., Elias, J., Burshtein, & V. R. Sharon, "OpenAI's GPT-4 performs to a high degree on board-style dermatology questions", *International Journal of Dermatology*, 63(1), 73-78, 2024.

[18] T. Talan, & Y. Kalınkara, "The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course", *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, 7(1), 33-40. 2023.

[19] J. Doughty, Z. Wan A. Bompelli, J. Qayum, T. Wang, J. Zhang, & M. Sakr, A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education, In Proceedings of the 26th Australasian Computing Education Conference, pp. 114-123. 2024.

[20] I. Azaiz, N. Kiesler, & S. Strickroth, Feedback-Generation for Programming Exercises With GPT-4 In Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, pp. 31-37. 2024.

[21] Y. Sonoda, R. Kurokawa, Y. Nakamura, J. Kanzawa, M. Kurokawa, Y. Ohizumi,... & O. Abe, "Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases", *Japanese Journal of Radiology*, pp. 1-5, 2024.

[22] M. Ishida, W. Gonoi, K. Nyunoya, H. Abe, G. Shirota, N. Okimoto,... & O. Abe, "Diagnostic Performance of GPT-4o and Claude 3 Opus in Determining Causes of Death From Medical Histories and Postmortem CT Findings", Cureus, 16(8), 2024.

[23] M. F. Firdaus, J. N. Wibawa, & F. F. Rahman, "Utilization of GPT-4 to Improve Education Quality Through Personalized Learning for Generation Z in Indonesia", IT for Society, 8(1), 2023.

[24] Madrid-García, A., Rosales-Rosado, Z., Freites-Nuñez, D., Pérez-Sancristóbal, I., Pato-Cour, E., Plasencia-Rodríguez, C., ... & Rodríguez-Rodríguez, L. "Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training", Scientific Reports, 13(1), 22129, 2023.

[25] M. G Rizzo, N. Cai, & D. Constantinescu, "The performance of ChatGPT on orthopaedic in-service training exams: A comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education", *Journal of Orthopaedics*, 50, 70-75, 2024.

[26] F. B. Warwas, & N. Heim, "Performance of GPT-4 in Oral and Maxillofacial Surgery Board Exams", Challenges in Specialized Question, 2024.

[27] J. Doughty, Z. Wan, A. Bompelli, J. Qayum, T. Wang, J. Zhang, & M. Sakr, A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. **In Proceedings of the 26th Australasian Computing Education Conference**, pp. 114-123, 2024.

[28] J. Samardžija, M. Žagar, & N. Drašković, "Work in progress: enhancing exam preparation with Chatgpt among university freshmen students." In ICERI 2024 Proceedings, pp. 9687-9694, 2024.

[29] V., Mavrych, P., Ganguly & O. Bolgova, "Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis.", Clinical Anatomy, 2024.

[30] M. Abu-Haifa, B. A. Etawi, H. Alkhatatbeh, & A. Ababneh, "Comparative Analysis of ChatGPT, GPT-4, and Microsoft Copilot Chatbots for GRE Test.", *International Journal of Learning, Teaching and Educational Research*, 23(6), 327-347, 2024.

[31] A. Koç & A. B. Öztiryaki, "Comparison of the accuracy performances of the Gemini Advanced, the GPT-4, the Copilot, and the GPT-3.5 models in medical imaging systems: A Zero-shot prompting analysis.", *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 13(4), 1216-1223, 2024.

[32] Öter, A., Ersöz, B., Bülbül, H. İ., & Sağıroğlu, Ş. (2024). Using Generative Artificial Intelligence in Exams: *A Research on KPSS with ChatGPT. International Journal of Educational Research Review*, 9(4), 269-274.

[33] C. Leiter, R. Zhang, Y. Chen, J. Belouadi, D. Larionov, V. Fresen & S. Eger, "Chatgpt: A meta-analysis after 2.5 months.", Machine Learning with Applications, 16, 100541, 2024.

[34] D. Milmo,. "ChatGPT reaches 100 million users two months after launch". The Guardian. ISSN 0261-3077, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app, 2024.

[35] S. Bulut., "Üretken Yapay Zeka: ChatGPT, Bing ve Bard Karşılaştırmalı Bir İnceleme", *International Journal Of Advanced Natural Sciences And Engineering Researches (Ijanser)* (7)9, 104 – 109. Doi: 10.59287/ijanser.1517. 2023.

[36] Internet: OpenAI, Hello GPT 4o, 2024. https://OpenAI.com/index/hello-gpt-4o/, 29.08.2024

[37] Internet:Gerttalkative. ChatGPT-4o vs GPT-4 vs GPT-3.5: What's the Difference?https://gettalkative.com/info/gpt-models-compared, 22.12.2024.

[38] Internet: S. Shubham, ChatGPT Statistics (AUG 2024) – Users Growth Data, 2024, https://www.demandsage.com/chatgpt-statistics/, 26.08.2024.

[39] Internet: Community, Update data Chat GPT fed with data until 09.2021, https://community.openai.com/t/update-data-chat-gpt-fed-with-data-until-09-2021/451130 , 26.08.2024.

[40] Internet: D. V. Meer , Number of ChatGPT Users and Key Stats (September 2024), https://www.namepepper.com/chatgpt-users, 29.08.2024

[41] Internet: Community, ChatGPT can now access the live Internet. Can the API? https://community.openai.com/t/chatgpt-can-now-access-the-live-internet-can-the-api/401928, 29.08.2024

[42] Internet: M. Diaz, R. Rajkumar, How to use Microsoft Copilot (formerly called Bing Chat), https://www.zdnet.com/article/how-to-use-Copilot/ 2024, 26.08.2024.

[43] Internet: Microsoft, Microsoft Copilot Prohttps://www.microsoft.com/en-us/store/b/Copilotpro, 29.08.2024.

[44] Internet: A. Subramanya, Gemini's big upgrade: Faster responses with 1.5 Flash, expanded access and more. https://blog.google/products/gemini/google-gemini-new-features-july-2024/, 27.08.2024.

[45] Internet: S. Pichai ve D. Hassabis, Introducing Gemini: our largest and most capable AI model, https://blog.google/technology/ai/google-gemini-ai/#capabilities, 27.08.2024.

[46] Internet: ÖSYM, Yükseköğretim Kurumları Sınavı, https://www.osym.gov.tr/TR,29434/2024-yuksekogretim-kurumlari-sinavi-2024-yks-temel-soru-kitapciklari-ve-cevap-anahtarlari-yayimlandi-09062024.html, 29.08.2024

[47] M. A. Cohen, "Some new evidence on the seriousness of crime.", Criminology, 26(2), 343-353, 1988.

[48] Internet: yks 2024'e ilk bakiş. https://www.linkedin.com/pulse/yks-2024e-ilk-bakiş-tedmem1-0mr5f/ , 29.08.2024

[49] Internet: M. Çakalp, Yapay zeka boğaziçini kazandı, https://www.hurriyet.com.tr/egitim/yapay-zeka-bogazicini-kazandi-42476894, 29.08.2024.

[50] Internet: G. Uyar, GPT-4, Üniversite Sınavına Girdi: Türkiye'nin %99'undan Daha Başarılı Oldu! https://www.webtekno.com/gpt-4-universite-sinavi-turkiye-yuzde-99-gecti-h133365.html, 21.12.204.

[51] G. Beutel, E. Geerits, & J. T. Kielstein, "Artificial hallucination: GPT on LSD?." Critical Care, 27(1), 148, 2023.

[52] T. B. Brown, "Language models are few-shot learners.", arXiv preprint arXiv:2005.14165, 2020.

[53] G. P. Reddy, Y. P. Kumar, & K. P. Prakash, Hallucinations in Large Language Models (LLMs)., In 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream) (pp. 1-6). IEEE, 2024

[54] J. Waldo, & S. Boussard, "GPTs and Hallucination: Why do large language models hallucinate?.", Queue, 22(4), 19-33, 2024.