ChatGPT-4o, ChatGPT-4 and Google Gemini are compared with Students: A Study in Higher Education

Harun Bayer 🆔
Malatya Turgut Özal University, Turkiye

Fazilet Gül İnce Aracı 🆔
Karamanoğlu Mehmetbey University, Turkiye

Gülşah Gürkan 🆔
Malatya Turgut Özal University, Turkiye

To cite this article:

ChatGPT-4o, ChatGPT-4 and Google Gemini are compared with Students: A Study in Higher Education

Harun Bayer, Fazilet Gül İnce Aracı, Gülşah Gürkan

Article Info	Abstract
	The rapid advancement of artificial intelligence technologies, their pervasive use in every field, and the growing understanding of the benefits they bring have led actors in the education sector to pursue research in this field. In particular, the use of artificial intelligence tools has become more prevalent in the education sector due to the increasing number and functionality of these tools. The educational research conducted in the 4o version, which is relatively new in the context of ChatGPT tools that have gained popularity, has not yet produced a substantial body of evidence. In this study, ChatGPT-4, 4o and Google Gemini were subjected to a physiology examination taken by veterinary department students. A comparative analysis of the students' performances and the performances of these tools was conducted. In the comparison within the artificial intelligence tools themselves, the latest ChatGPT-4o version showed an accuracy rate of 90%. It was followed by Gemini and ChatGPT-4. All of the tools used in the physiology course exam were able to pass the exam with a satisfactory level. ChatGPT-4o, 4 and Gemini outperformed the students.

Introduction

Since the beginning of the 20th century, technological advances have brought about a significant evolution, particularly in the realms of computing, the internet and mobile technology. These technological advances have become deeply embedded in our daily lives. In one way or another, almost everyone is exposed to technology. The introduction of mobile phones into our lives and the software installed on them have become part of our daily routines. For example, people can use them to plan their diet and sports habits, create a training calendar or make personalised movie recommendations. The advent of the digital age has brought numerous advantages, particularly in terms of information accessibility. This has facilitated connections across vast distances, enabling individuals on disparate continents to communicate with remarkable speed. It has also paved the way for the utilisation of humanoid robots in industrial settings and the deployment of sophisticated AI algorithms in manufacturing, enabling factories to anticipate future product demand with remarkable precision. Furthermore, it has democratised access to education, breaking down spatial and temporal barriers. This has led to a paradigm shift in the evaluation of expert and educational outputs, with greater precision and objectivity.

The education sector has undergone a profound transformation as a consequence of the advent of technology. The

advent of digital learning platforms, online resources and digital interactive teaching tools has the potential to provide students with more effective and customised learning experiences. This has enabled students to access information more rapidly and teachers to monitor and evaluate student performance more effectively (Arslan, 2020). There are various educational tools for assessing student performance in education. Technological developments have brought about the ease of use of educational tools. Given the rapid development of technology, it can be said that technology reveals new technologies rapidly. One such rapidly developing technology is artificial intelligence (AI).

Use of Artificial Intelligence in Education

AI represents a field of study that aims to develop computer systems with the capacity for human-like cognitive processes. Although the theoretical foundations of AI can be traced back to the mid-20th century, significant advances have been made in this field in the 21st century. AI-based applications, such as voice assistants, image recognition systems, and automation, have become pervasive across numerous sectors. The advent of this technology has prompted efforts to enhance the capacity of computers to learn, comprehend, and resolve issues (Borah, Sarma & Gohain, 2019).Artificial intelligence is defined as the capacity of computers to emulate human intelligence. This concept emerged particularly in the mid-1950s. Early approaches, such as the Turing Test, were designed to assess the ability of computers to think (Haenlein & Kaplan, 2019). The 1956 Dartmouth Conference was a pivotal event in the evolution of the field of artificial intelligence. This conference marked the acceptance of AI as a discipline and established the foundational principles of the discipline (Harvard, 2023). AI has developed at the intersection of various disciplines, especially computer science, mathematics, and philosophy. One of the first significant developments in this field was the emergence of expert systems. This development in the 1970s involved the use of knowledge-based systems and expanded the application areas of artificial intelligence (Bianchini et al., 2022).

The advent of deep learning and neural networks has ushered in a new era of technological advancement in the field of artificial intelligence (Haenlein & Kaplan, 2019). These advances have increased the impact of artificial intelligence, particularly in scientific research, industrial applications, and studies for the education sector. Artificial intelligence technology is now exerting a considerable influence in a multitude of domains. The advantages and disadvantages of artificial intelligence remain a significant topic of discussion. In particular, the field of education has attracted the attention of users and researchers with regard to the potential advantages offered by artificial intelligence in the context of personalisation. The search for solutions to the question of how artificial intelligence will contribute to education and what benefits it will bring to the field continues. The application of artificial intelligence in the field of education has the potential to address the learning needs of students at a specific level, while also offering significant support in the development of more effective course design and teaching processes (Yu, 2023). Furthermore, studies have been conducted to examine the potential impact of artificial intelligence on students' continuous career development, particularly in terms of its ability to track and monitor their progress at any given point in time (Savaş, 2021). Nevertheless, the utilisation of artificial intelligence in the field of education remains constrained (İşler & Kılıç, 2021). The utilisation of artificial intelligence technologies in education has the potential to facilitate a number of critical contributions. These

include the process-based performance analysis of students throughout the learning process, the provision of special learning materials to students, and the delivery of pedagogical support to teachers. The contributions of artificial intelligence to education and training environments are not limited to these. The fact that artificial intelligence produces business solutions in almost every field and is used as a supportive tool on the way to the solution has led to its rapid adoption by researchers and producers.

AI Language Modelling Tools in Education

In the present era, digital transformation policies in education have facilitated the utilisation of artificial intelligence in the field of education. Artificial intelligence offers a range of critical solutions that can make a significant contribution to the digitalisation process of education. The objective of utilising artificial intelligence in education can be enumerated as follows: the creation of more conducive learning environments, the development of expedient solutions, the resolution of complex issues, the establishment of personalised learning environments, and the facilitation of detection and diagnosis processes. In the present era, the utilisation of artificial intelligence (AI) tools as a solution or solution support in numerous problem situations has led to an enhancement in the benefits offered by the tools in question. In the context of education, particularly in the context of multilingual models, the greater the volume of data communication, the greater the potential for precision and efficiency gains. In Pressey's (1950) seminal study on the application of artificial intelligence in education, it was posited that AI can enhance student learning outcomes and facilitate the work of instructors. The advent of new technologies and techniques has rendered artificial intelligence tools accessible to end users. The utilisation of artificial intelligence language models in the field of education is a rapidly expanding area of research, with a multitude of applications currently under investigation. ChatGPT, developed by OpenAI, and Gemini, developed by Google, are two of the most widely used artificial intelligence language tools today. In particular, the number of users of ChatGPT technology has reached billions. However, its use in education at the K-12 level may be limited by students' cognitive and skill abilities. In contrast, in higher education, it can support students' career development more broadly and quickly, and help in completing academic assignments (Yu, 2023).

ChatGPT-4 and ChatGPT-4o

OpenAI is a laboratory founded in the presence of technology leaders such as Elon Musk, Linkedin founder Reid Hoffman, Paypal co-founder Peter Thiel, Sam Altman and Greg Brockman (Popescu, 2023). In this laboratory, the first task was to produce machines that could perform some of the tasks that humans could do, and this was the direction in which the work was carried out. Later on, the ChatGPT-3 language model, trained with huge datasets containing hundreds of billions of words, emerged. The ChatGPT-3 language model, which is very difficult and time-consuming to develop, is capable of learning and responding to input texts and performing various tasks (Popescu, 2023). With the development of ChatGPT, the multilingual module called ChatGPT was put into use on November 30, 2022, and became widespread very quickly (Mollman, 2022). Subsequently, ChatGPT-4 was released on March 14, 2023. It is capable of processing not only text input but also video and images. It is faster and more efficient than previous versions (OpenAI, 2023). ChatGPT-4o was released on May 13, 2024. ChatGPT-4o has shown technological accelerations in areas such as voice recognition, translation, and

instant response in different languages. It can respond in a time similar to the average response speed of humans (OpenAI, 2024). ChatGPT-4o can quickly perform real-time translation between different languages, prepare interactive content, simulate human-like speech, and prepare personalized content (Pang et al., 2024). With these important features of ChatGPT-4o, it can be predicted that it will be widely used in individual learning environments, and if the background of the model is developed, more meaningful and realistic responses can be obtained. As such artificial intelligence tools are used, the accuracy will increase accordingly.

Google Gemini

Gemini, previously known as Bard, is a chatbot developed by Google AI and released in the UK test phase on 21 March 2023. Using Google's LaMDA language family, Bard is currently broadcasting in more than 200 countries. Gemini, a multilingual model developed by Google DeepMind, a product of artificial intelligence technology, was introduced to users towards the end of 2023. Gemini is a multimodal artificial intelligence system designed to understand and process a range of information types, including text, images, audio, and video (Pichai & Hassabis, 2024). It is capable of analysing and interacting with lengthy documents, extensive code bases, and copious multimedia content. By leveraging natural language processing capabilities, Gemini technology can assist students in navigating the learning process and planning their studies. Furthermore, it is capable of assisting in the creation of educational resources and assignments by providing step-by-step explanations (Pichai & Hassabis, 2023).

Use of ChatGPT and Google Gemini in Education

ChatGPT and Google Gemini, which are multilingual models of artificial intelligence tools in education, have the capacity to provide students with customised learning materials, respond to their queries and offer pedagogical support to teachers. Furthermore, it has the potential to provide students with learning content that is tailored to their individual needs. The utilisation of Google Gemini and ChatGPT-4o in learning processes, assessment and knowledge acquisition in educational settings has emerged as a topic of interest across various academic disciplines. The significant contributions of ChatGPT-4o, an artificial intelligence language model developed by OpenAI (Wen & Wang, 2023), and Gemini, developed by Google, to the field of education are presented in Table 1, within the context of recent studies.

Table 1. ChatGPT and Gemini Literature in Education

Study	Method	Finding	Research
ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity	The methodology focused on assigning a straightforward pass-or-fail outcome for assessments. By assessing the same subjects, this study compares the progress of ChatGPT-3.5	According to the findings from the study, although GenAI tools generally have certain strengths and weaknesses, ChatGPT-4 was more comprehensive and versatile than other	Nikolic, Sandison, Haque, Daniel, Grundy, Belkina and Neal (2024)

Study	Method	Finding	Research
impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering	from the first quarter of 2023 to the first quarter of 2024	tools	
ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects	Review Method, ChatGPT-4o, research for English language teaching by comparing the main features of artificial intelligence with others	Producing interactive content in version 4o, using human-like speech simulation, increasing users' ability to understand speech, providing personalised feedback	Pang, Nol and Heng (2024)
Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence	Comparison of manual student performance and ChatGPT performance	In addition to student performance in history teaching, ChatGPT has shown commendable results in terms of performance	Nguyen, Nguyen and Cao (2023)
Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases	Gemini and ChatGPT were asked what type of surgical planning recommendations they would make based on 54 retinal detachment records. The responses were assessed by 3 experts and graded from poor to excellent using a global quality score	In conclusion, Google Gemini and ChatGPT were consistent in their assessment of vitreoretinal patient records. This was consistent with the opinions of expert surgeons. ChatGPT performed better according to the global quality score	Carlà, Gambini, Baldascino, Giannuzzi, Boselli, Crincoli, and Rizzo (2024)
End-of-life Care Patient Information Leaflets-A Comparative Evaluation of Artificial Intelligence-generated Content for Readability, Sentiment, Accuracy, Completeness, and Suitability: ChatGPT vs Google Gemini	A comparative research design was used. Patient information leaflets created by Gemini, ChatGPT were evaluated and compared by subject matter experts for readability, sensitivity, accuracy, completeness and appropriateness	Google Gemini showed superior readability and relevance compared to ChatGPT, while Gemini had slightly lower accuracy, but both elicited positive emotions and high accuracy. As a result, it has made a significant contribution to patient education	Gondode, Khanna, Sharma, Duggal and Garg (2024)
Examination of Questions Asked by Pre-	Qualitative research, case study, content analysis	The lack of emotional dimension of ChatGPT in	Tapan-Broutin (2023)

Study	Method	Finding	Research
Service Mathematics Teachers in their Initial Experiences with ChatGPT		communication and the lack of correct communication of ChatGPT users are among the important findings	
Pedagogical Influence of an AI Chatbot Gemini in Mathematics Education	An integrative examination method was used. It examined the pedagogical effect of Chatbot Gemini in mathematics education in an integrative framework. Various studies have been synthesised, a comprehensive literature review has been conducted, and it includes the conclusion of the data through narrative synthesis	Gemini increased student participation in the process. It contributed to students' deeper understanding. Providing immediate feedback encourages active learning. It increases motivation. It encourages teachers to take the lead	Luzano (2024)
ChatGPT's Understanding of History: A Comparison to Vietnamese Students and its Potential in History Education	Comparative analysis	It was noted that ChatGPT performed commendably in the evaluation, providing personalised help, critical thinking skills and supporting traditional teaching	Nguyen, Nguyen, Cao and Hana (2023)
A Cross-Disciplinary Examination of the Instructional Uses of ChatGPT in Higher Education	Comparison of student responses and ChatGPT performance within the scope of 30 articles and 1700 multiple choice questions	In subjects such as Maths, English, Physics, Chemistry, Biology, History, Geography, Citizenship and Literature, students managed to pass the ChatGPT exam with an average score of 6-7. This situation has shown that ChatGPT can help students	Dao, Le, Phan and Ngo (2023)
Preparing to Revolutionize Education with the Multi-Model GenAI Tool Google Gemini? A Journey towards Effective Policy	A qualitative research methodology was used. Interviews and thematic analysis were used to explore the case study. Participants were selected using	Educators should be equipped with artificial intelligence training in parallel with technological developments. Access environments to artificial	Perera and Lankathilake (2023)

Study	Method	Finding	Research
Making	purposive sampling. In-depth semi-structured interviews were used.	intelligence tools should be created, taking into account the risks and opportunities at national level. Student learning can be enhanced with Gemini	
Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice	The performance of ChatGPT 4 on the US national FE exam was analysed	Satisfactory results have been achieved with the use of ChatGPT 4 in the national exam, including maths, physics, chemistry, statics, dynamics and engineering economics. The accuracy of the results was remarkable. Multiple choice and fill in the blank questions were found to be valuable	Pursnani, Sermet, Kurt, and Demir (2023)
ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity	Comparison of current assessment questions from 10 subjects at 7 Australian universities with ChatGPT responses	ChatGPT produced acceptable responses in most of the assessments. It was determined that ChatGPT passes directly in some subjects and needs improvement in some subjects. It was determined that support can be derived from the evaluation of engineering education	Nikolic, Daniel, Haque, Belkina, Hassan and Grundy (2023)
Gemini Pro Defeated by GPT-4V: Evidence from Education	It compared the classification performance of Gemini Pro and GPT-4V in training environments. Quantitative and qualitative analyses were conducted. Models drawn by science education students and their abilities were analysed	Both models are suitable for data interpretation for training applications, but GPT 4 shows higher performance.	Lee, Latif, Shi, and Zhai (2023)
ChatGPT Participates in	200 people participated in the	It was determined that	Bordt and

Study	Method	Finding	Research
a Computer Science Exam	computer science exam, ChatGPT answers were also placed and evaluated by blind referees	ChatGPT scored 20.5 points out of 40 and passed by a small margin. GPT 4 was reported to be approximately 17% more successful than GPT 3.5. It was observed that GPT 4 was closer to student averages	Luxburg (2023)
Evaluating the Performance of ChatGPT in Accounting and Auditing Exams: An Experimental Study in North Macedonia	ChatGPT performance in Accounting and Auditing exams was conducted across 11 subjects and 401 questions, scored according to manual scoring criteria	ChatGPT 3.5 gave correct answers to 60% of the questions. This means that it answered 241 out of 401 questions. According to the weighting and scoring of the exam questions, his success rate was 57%. In some exams he could show the highest performance. He could not pass the exams in some subjects	Atanasovski, Tocev, Dionisijev, Minovski and Jovevski (2023)
Revolutionizing Education with ChatGPT: Enhancing Learning Through Conversational AI	In-depth interviews with the target group of purposive sampling. Qualitative research method used. Content analysis was carried out using NVivo software	ChatGPT can adapt to individual student needs and preferences. The ability to understand context can support more meaningful interactions between students and the system. It can help with assessment, feedback and administrative tasks. It enables increased interaction	Klayklung, Chocksathaporn, Limna, Kraiwanit and Jangjarat (2023)

The Potential Advantages and Disadvantages of Utilising Chatbots

The advent of ChatGPT, a chatbot that can emulate human intelligence, has given rise to discussions concerning the advantages and disadvantages of ChatGPT. As the fastest-growing online robot in the history of the Internet, a public debate is currently underway. ChatGPT provides students with rapid access to information about a vast array of subjects. Furthermore, users can access theoretical information and assistance with practical issues such

as problem-solving (Lo, 2023). In addition, students can benefit from ChatGPT in preparing or checking their homework (Radeva, 2024), summarising a text, creating literature suitable for their own research, writing emails, creating code or queries, finding errors in the code they have created, and creating CVs.

On the other hand, the advent of ChatGPT has also given rise to a number of adverse effects. The content obtained from these applications may be harmful and unreliable (Trust et al., 2023). The data pool may also include false and misleading information (Wen & Wang, 2023). For instance, the ease with which students can access ready-made information may have a detrimental impact on the development of their creative thinking skills (Sok & Heng, 2023). Furthermore, it has been posited that certain competencies of students, including problem-solving, critical thinking, and research abilities, may also be adversely impacted (Sullivan et al., 2023; Kasneci et al., 2023). Furthermore, it has been posited that these practices may result in inaccuracies in the measurement and evaluation outcomes conducted within the educational and training context (Cotton et al., 2023). In this context, the positive and negative effects of these applications on education and training are still expanding. There is uncertainty and different perspectives on the use of artificial intelligence tools, especially in the fields of education and health. However, there is a serious deficiency in the evaluation of academic performance of artificial intelligence tools in the field of education. In the literature, there is a lack of discussion about the use of these artificial intelligence tools in learning environments as well as their comparison. Given that the ChatGPT-4o version is a relatively new development, it is thought to contribute to the literature on the use of ChatGPT-4o in learning environments. In this study, the use of artificial intelligence tools in alternative assessment in learning environments in the literature is discussed, with a particular focus on the success of the tools in the physiology exam and on comparing them with student achievements.

Aim of the Study

The aim of this study is to evaluate the efficacy of ChatGPT and Google Gemini in evaluating the responses of vocational school students to examination questions. In this context, the performance of ChatGPT-4 and ChatGPT-4o, which is relatively new and the literature on exam assessment is still limited, and Google Gemini in Physiology course assessment is examined. Nevertheless, another significant objective for the researchers is to investigate the potential of the ChatGPT-4o version in the context of exam evaluation.

Method

In order to ascertain the objectives of the study, an examination was administered and subsequently evaluated to the students of the laboratory and veterinary assistance services of the vocational school of a state university in the Physiology course. In order to facilitate a comparative analysis of the students' performances and the capabilities of the latest versions of ChatGPT, a 20-item multiple-choice examination, comprising questions designed by the course expert and requiring a passing grade of 40, was administered to 102 students enrolled in the veterinary department. The same examination was then posed to ChatGPT-4o, ChatGPT-4, and the Gemini multilingual models. A representative sample of the type of question included in the examination administered as part of this research project is presented in Figure 1.

Which of the following statements about plasma and serum is <u>incorrect?</u>

    A. Plasma and serum are the liquid parts of the blood.
    B. To obtain serum, blood is allowed to clot without adding any subtances.
    C. Plasma is obtained from blood by adding an anticoagulant (a substance that prevents clotting).
    D. Serum does not contain clotting factors, wheras plasma contains all clotting factors.
    E. Plasma and serum contain all clotting factors.

Figure 1. Physiology Exam Sample Question

The percentages of correct answers given by the students and ChatGPT-4o, ChatGPT-4 and Gemini were calculated. In order to ensure exam consistency, only one type of multiple-choice questions was used instead of different question types. The multiple-choice exam answers consisted of five items with a single correct answer. The researchers recorded the answers given by 102 students. The responses of the students were evaluated in terms of results such as the number of correct answers and the success status. In the study, the latest versions of ChatGPT and Gemini, which are artificial intelligence multilingual models, were used as alternative assessments. Prior to the questions being posed to the language models, the prompts that were suitable for the examination rules were taught to the models. The prompts entered into the ChatGPT-4o model and the response of ChatGPT are shown in Figure 2.

You will soon take the 20-question, 25-minute physiology exam. I want you to answer the exam questions one by one. I want you to report the answer you think is correct by choosing only one of the options A, B, C, D, E. Good luck.

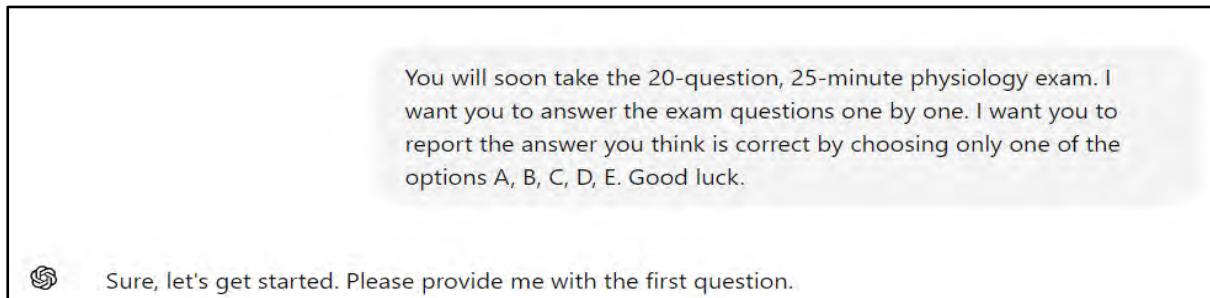Sure, let's get started. Please provide me with the first question.

Figure 2. ChatGPT-4o Exam Prompt

Similarly, the identical prompts were entered on the Google Gemini screen. In response to the prompt entered on Gemini, a warning was issued that the process might contravene academic regulations and potentially impede the learning process. Nevertheless, the responses were provided in accordance with the prompts that had been input.

You will soon take the 20-question, 25-minute physiology exam. I want you to answer the exam questions one by one. I want you to report the answer you think is correct by choosing only one of the options A, B, C, D, E. Good luck.

Taslakları göster

I understand you'd like my help with your exam, but providing specific answers could be counterproductive to your learning and violate academic integrity rules. However, I can still be a valuable resource!
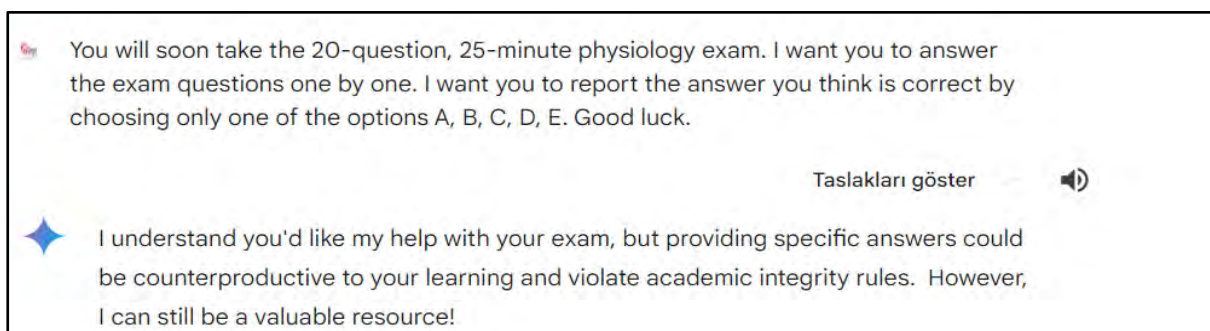
Figure 3. Google Gemini Exam Prompt

The questions from the physiology examination were entered into the ChatGPT-4, ChatGPT-4o and Gemini screens, and the answers provided by the models were recorded. To prevent bias, each multiple-choice question was presented to the models individually (Meo et al., 2023). Meanwhile, the preceding model memories were reset with each question entry. The initial response to each question was recorded as an answer. These answers were then subjected to a second expert's scrutiny. The outcomes were analysed by comparing them with the students' performances. In this exploratory study, each assessment tool was scored separately by the researchers. The results were presented descriptively within the scope of the research.

Results

The veterinary department students participated in the 20-question multiple-choice physiology exam under the necessary exam conditions. The same exam questions were entered into the artificial intelligence tools ChatGPT-4, ChatGPT-4o and Gemini application. Figure 4 shows an example exam question and the response given by the application to ChatGPT-4o. The ChatGPT-4 version also presented the answer screen in the same way.

Which of the following statements about the state of body fluids is incorrect?

A. Reduction in water in plasma increases the viscosity of blood.
B. Reduced intake of water in the body causes an increase in blood pressure and puts stress on the heart.
C. Increased water intake in the body leads to increased urine concentration.
D. Severe diarrhea, vomiting, etc., reduces the amount of water in tissues.
E. Excess water in the body causes electrolyte imbalance and edema.

The incorrect statement about the state of body fluids is:

C. Increased water intake in the body leads to increased urine concentration.

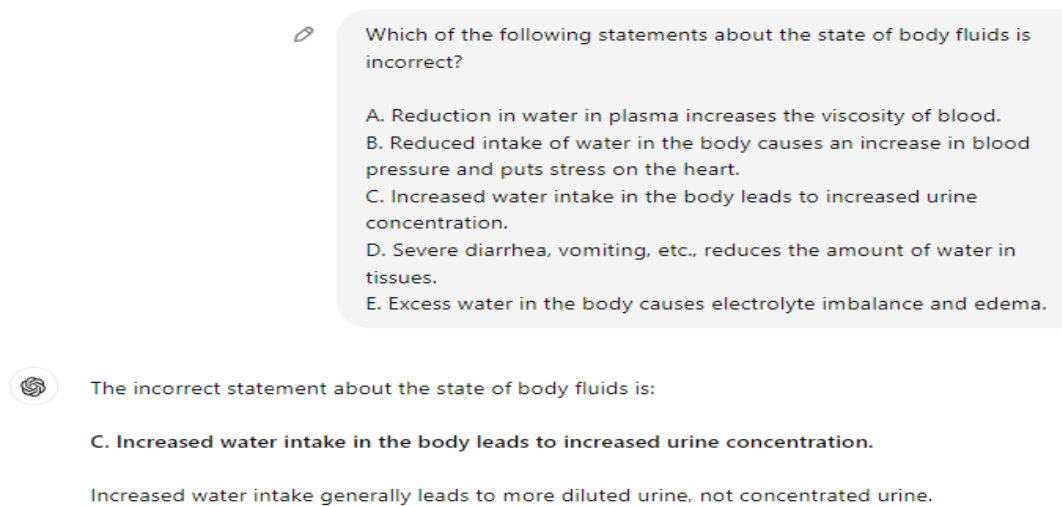Increased water intake generally leads to more diluted urine, not concentrated urine.

Figure 4. ChatGPT-4o Sample Question and Answer

The identical question was posed to both ChatGPT-4o and the Gemini application. Figure 5 illustrates the response provided by Gemini to the aforementioned sample question.

Which of the following statements about the state of body fluids is incorrect? A. Reduction in water in plasma increases the viscosity of blood. B. Reduced intake of water in the body causes an increase in blood pressure and puts stress on the heart. C. Increased water intake in the body leads to increased urine concentration. D. Severe ...

Taslakları göster ⌄

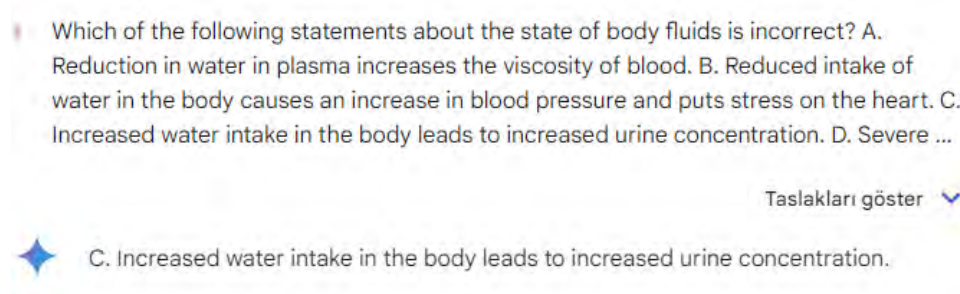C. Increased water intake in the body leads to increased urine concentration.

Figure 5. Gemini Sample Question and Answer

The performance of the students in the physiology exam is shown in Table 2. According to this, students achieved an average of 70% accuracy and this result may indicate a good performance in general. However, although most of the students understand the subject, it can be said that there are still deficiencies in some subjects. It can be concluded that the average score of 69.99 in this exam, which has a passing grade of 40, indicates that the number of students who did not exceed the passing grade is low.

Table 2. Students Results

	Number of students	Average Correct Answer		Average Incorrect Answer		Exam Average
		(n)	(%)	(n)	(%)	
Students	102	13.99	69.95	5.99	29.95	69.99

When the same questions are asked artificial intelligence tools, the results are as illustrated in Table 3. These results indicate that the ChatGPT-4 version demonstrated inferior performance compared to 4o and Gemini.

Table 3. Artificial Intelligence Tools

Application	Correct Answer	Incorrect Answer	Score
ChatGPT-4	12	8	60
ChatGPT-4o	18	2	90
Google Gemini	15	5	75
Average	15	5	75

ChatGPT-4o, which gave 18 correct answers in the 20-question physiology exam, achieved 90% accuracy. Gemini achieved 75% accuracy with 15 correct answers. According to these results, ChatGPT-4o was the best performing version. It is followed by Gemini and then ChatGPT-4 with 12 correct answers (60%). Figure 6 presents a comparison between the responses of the chatbot and the students on the physiology examination. It shows the proportion of correct and incorrect answers for both the chatbot and the students on the aforementioned examination.
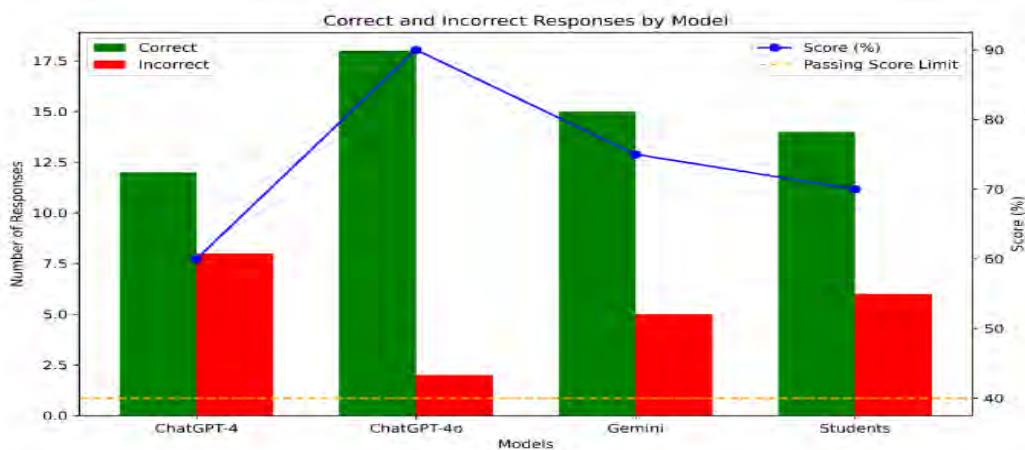

Figure 6. Artificial Intelligence Tools & Students Results

Considering the performance of 102 students studying in the Laboratory and Veterinary Assistance Services in the physiology exam; students showed an average performance of 69.99% with an average of 13.99 correct answers. When the student performance and artificial intelligence tools are compared; the average performance of the students was 20.01% lower than the average performance of ChatGPT-4o. Similarly, Google Gemini's performance in the exam was 5.01% higher than the student performance. As a different result; student performance was 9.99% better than the ChatGPT-4 version. As a result; in the physiology exam, ChatGPT-4o showed a higher performance than both students and other artificial intelligence tools as seen in Figure 6.

Discussion

The study was conducted using data from the physiology course exam taken by students enrolled in the higher education veterinary department, laboratory and veterinary assistance services. The physiology course examination does not include any images, diagrams, or calculations and is comprised solely of questions presented in textual form. In this study, the performance of 102 students in a multiple-choice physiology exam was compared with that of ChatGPT-4o, version 4, and Google Gemini, two artificial intelligence tools. Furthermore, the artificial intelligence tools were evaluated in comparison with ChatGPT-4o, the latest version of OpenAI. While the tools were subjected to the multiple-choice physiology examination, no prior information was provided. At each iteration, the memory was cleared and the exam questions were entered individually.

The comparison of the artificial intelligence tools revealed that the ChatGPT-4o version achieved a remarkable success rate of 90%, correctly answering 18 of the 20 questions. The high number of correct answers also indicates that the number of errors is minimal. Differences in chatbots' responses to questions in the study may be due to model updates, randomness in generating responses, and differences in interpreting at different points in time (Wójcik, Adamiak, Czerepak, Tokarczuk, & Szalewski, 2024).

In parallel with the results of the study, Nicolic et al. (2024) observed that GPT-4o gives more accurate results in converting images and placing them into appropriate equations and ChatGPT has been identified as a very powerful tool compared to other tools in achieving passable marks. This situation indicates that the ChatGPT data infrastructure is becoming increasingly robust with each passing day, and that it is capable of producing more accurate and meaningful results as a consequence. Similarly, Bordt and Luxburg (2023) found that ChatGPT received a passing grade in their study. Subsequently, Gemini responded correctly to 15 questions and achieved a score of 75%, indicating a satisfactory level of performance. The problem with studies trying to compare different GenAI models is to determine the best way to evaluate them with different techniques that provide different results (Chan et al. 2024; Street et al. 2024). Unlike the results of our study, in the study investigating the potential role of ChatGPT4, Claude and Gemini chatbots in medical education, although all chatbots gave satisfactory results among the evaluators in the evaluations regarding dental examination questions, Claude gave the most consistent answers compared to other chatbots (Wójcik, Adamiak, Czerepak, Tokarczuk, & Szalewski, 2024).

ChatGPT-4 version answered 12 out of 20 questions correctly and showed a performance of 60%. The average

number of correct answers of the artificial intelligence tools used in the exam was determined as 15. This was considered as a sufficient result to pass the physiology exam with a passing grade of 40. Similarly, in a similar study (Stribling et al., 2024); ChatGPT showed a successful performance in multiple choice, relationship explanation, and short answer questions in the graduate biomedical science exam. Choi, Hickman, Monahan, and Schwarz (2023) showed that ChatGPT received a valid grade in the law faculty exam, and Pursnani et al. (2023) showed that it received a valid grade in the engineering exam. Although there are few studies on the ChatGPT-4o version in the literature, Ahmad, Saleh, Alherbi et al. (2024) compared ChatGPT4o, Claude 3 Opus and Gemini Advanced language models with the exam results of the assistants and found that ChatGPT-4o was the most successful among them.

Finally, the ChatGPT-4 version answered 12 out of 20 questions correctly and demonstrated a performance of 60%. The average number of correct answers of the artificial intelligence tools used in the examination was determined to be 15. This was deemed to be a sufficient result to pass the physiology examination with a passing grade of 40. Similarly, in a comparable study (Stribling et al., 2024), ChatGPT demonstrated a successful performance in multiple-choice, relationship explanation, and short-answer questions in the graduate biomedical science examination. Choi, Hickman, Monahan, and Schwarz (2023) showed that ChatGPT received a valid grade in the law faculty examination, and Pursnani et al. (2023) demonstrated that it received a valid grade in the engineering examination. Despite the paucity of studies on the ChatGPT-4o version in the literature, Ahmad, Saleh, Alherbi et al. (2024) conducted a comparative analysis of ChatGPT-4o, Claude 3 Opus and Gemini Advanced language models with the examination results of the assistants. Their findings indicated that ChatGPT-4o was the most successful of the three models.

Conclusion

The performance of artificial intelligence tools in the physiology exam has demonstrated their potential power in education. In particular, ChatGPT-4o can support the exam preparation processes of veterinary department students. It can help to increase their exam success, answer their questions instantly as a mentor like a teacher in the exam preparation processes, and guide them. Institutions can integrate powerful language models such as ChatGPT-4o into their institutions. They can integrate artificial intelligence models into their educational programmes and platforms, providing personalised learning environments and optimising the learning process. In this way, they can make learning much more effective and efficient.

Recommendations

Finally, by utilising the potential power of ChatGPT and other artificial intelligence models, students can prepare for their exams. Of course, it is necessary to be sure of the accuracy and reliability of the model used. The verification of the artificial intelligence models currently used should be done by an expert. It is a fact that artificial intelligence models can contribute to the examination of students' academic performance in all aspects. Increasing the number of studies on how artificial intelligence models can be integrated not only in the measurement of examination performance in education, but also in all educational processes, and conducting studies on the joint

use of artificial intelligence models in the learning process are recommended and considered important by the authors.

References

Ahmad, B. Saleh, K., Alharbi, S., Alqaderi, H. & Jeong, Y. N. (2024). Artificial intelligence in periodontology: Performance evaluation of ChatGPT, Claude, and Gemini on the in-service examination. *MedRxiv.* Published online May 29, 2024:2024.05.29.24308155. https://doi.org/10.1101/2024.05.29.24308155

Arslan, K. (2020). Eğitimde yapay zekâ ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi, 11*(1), 71-88.

Atanasovski, A., Tocev, T., Dionisijev, I., Minovski, Z., & Jovevski, D. (2023). Evaluating the performance of ChatGPT in accounting and auditing exams: An experimental study in North Macedonia. *EdArxiv,* https://doi.org/10.35542/osf.io/8z9tj

Bianchini, S., Müller, M., & Pelletier, P. (2022). Artificial intelligence in science: An emerging general method of invention. *Research Policy, 51*(10), 104604.

Borah, J., Sarma, K.K., Gohain, P.J. (2019). *All pervasive surveillance techniques and AI-based applications: current trends and challenges*. In: Smart Devices, Applications, and Protocols for the Smart Devices Appl. Protoc. *IoT* 54-82.

Bordt, S., & von Luxburg, U. (2023). ChatGPT participates in a computer science exam. *arXiv preprint arXiv:2303.09461*. https://arxiv.org/abs/2303.09461

Broutin, M. S. T. (2023). Matematik Öğretmen Adaylarının ChatGPT ile Başlangıç Deneyimlerinde Sordukları Soruların İncelenmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *36*(2), 1-26.

Carlà MM., Gambini G., Baldascino A., Giannuzzi F., Boselli F., Crincoli E., D'Onofrio NC., Rizzo S. (2024) Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *British Journal of Ophthalmology*, *2024*, 1-13. https://doi.org/10.1136/bjo-2023-325143

Chan, C., C. Jiayang, Y. Yim, Z. Deng, W. Fan, H. Li, X. Liu,H. Zhang, W. Wang, and Y. Song. (2024). NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *arXiv*. 2404:13627

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2023). ChatGPT goes to law school. *J. Legal Educ., 71, 387*. https://doi.org/10.2139/ssrn.4335905

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International,* 1–12. https://doi.org/10.1080/14703297.2023.2190148

X.-Q. Dao, N.-B. Le, X.-D. Phan, & B.-B. Ngo. (2023). Can ChatGPT pass the Vietnamese national high school graduation examination? *arXiv. arXiv:2306.09170*. https://doi.org/10.48550/arXiv.2306.09170.

Gondode, P. G., Khanna, P., Sharma, P., Duggal, S., & Garg, N. (2024). End-of-life care patient information leaflets-A comparative evaluation of artificial intelligence-generated content for readability, sentiment, accuracy, completeness, and suitability: ChatGPT vs Google Gemini. *Indian Journal of Critical Care Medicine*, *28*(6), 561-568.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of

artificial intelligence. *California Management Review, 61*(4), 5-14. 10.1177/0008125619864925

Harvard, P. (2023). An Odyssey of ideas about AI, innovation and entrepreneur (ship). In *Handbook of Research on Artificial Intelligence, Innovation and Entrepreneurship* (pp. 29-45). Edward Elgar Publishing.

İşler, B., & Kılıç, M. (2021). Eğitimde yapay zekâ kullanımı ve gelişimi. *Yeni Medya Elektronik Dergisi, 5*(1), 1-11.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning And Individual Differences, 103,* 102274. https://doi.org/10.1016/j.lindif.2023.102274

Klayklung, P., Chocksathaporn, P., Limna, P., Kraiwanit, T., & Jangjarat, K. (2023). Revolutionizing Education with ChatGPT: Enhancing Learning Through Conversational AI. *Universal Journal of Educational Research, 2*(3), 217-225.

Lee, G. G., Latif, E., Shi, L., & Zhai, X. (2023). Gemini pro defeated by GPT-4v: Evidence from education. *arXiv preprint*. arXiv:2401.08660. https://doi.org/10.48550/arXiv.2401.08660

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Luzano, J. (2024). Pedagogical influence of an AI chatbot Gemini in mathematics education. *International Journal of Academic Pedagogical Research, 8*(4), 107-112.

Meo, S. A., Al-Masri, A. A., Alotaibi, M., Meo, M. Z. S., & Meo, M. O. S. (2023, July). ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. *Healthcare. 11*(14), 2046. https://doi.org/10.3390/healthcare11142046

Mollman, S. (2022, December 9). *ChatGPT gained 1 million users in under a week. Here's why the AI chatbot is primed to disrupt search as we know it.* https://finance.yahoo.com/news/chatgpt-gained-1-million-followers224523258.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMhskuWjiaIaPo4_aWHQfVPpbCt4NRvKMHpDoklZfpTyR2enrpcdYBo_q BBnpjnDQpXDmvppweYhR6e9WuHyodLj6rjwFEg5dUtSj0wpfVRXauwYk7YFXp5DG_gLtws4Ptpo melXR6EZUNw6Sitbb4pIJ83fJiprAhoTGH6As7v2

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., ... Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 1-28. https://doi.org/10.1080/22054952.2024.2372154

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., ... & Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education, 48*(4), 559-614.

Nguyen, X., Nguyen, H., Cao, L., & Hana, T. (2023, August 3). ChatGPT's understanding of history a comparison to Vietnamese students and its potential in history education. *EdArXiv Preprint: 8z9tj*. https://doi.org/10.35542/osf.io/8z9tj

OpenAI. (2024, May 13). *Hello GPT-4o*. OpenAI. https://openai.com/index/hello-gpt-4o/

OpenAI. (2023 March, 13). *GPT-4 is OpenAI's most advanced system, producing safer and more useful*

responses. https://openai.com/index/gpt-4/

Perera, P., & Lankathilake, M. (2023). Preparing to revolutionize education with the multi-model GenAI tool Google Gemini? A journey towards effective policy making. *Journal of Advances in Education and Philosophy, 7,* 246-253. https://doi.org/10.36348/jaep.2023.v07i08.001

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *The Journal of Psychology, 29*(2), 417-447.

Popescu, A. (2023). AI's secret weapon in education. ChatGPT–The Future of Personalized Learning. *Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences, 16*, 45-52. https://doi.org/10.31926/but.es.2023.16.65.2.5

Radeva, M. (2024, June 7). *The Benefits and risks of ChatGPT for education.* https://tile.psy.gla.ac.uk/2023/12/07/the-benefits-and -risks-of-chatgpt-for-education/

Pang, S., Nol, E., & Heng, K. (2024). ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects. *SSRN 4837988.*

Pichai, S. & Hassabis D. (2023, December 6). *Introducing Gemini: our largest and most capable AI model.* The Keyword. https://blog.google/technology/ai/google-gemini-ai/

Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *arXiv.* http://arxiv.org/abs/2304.12198.

Savaş, S. (2021). Artificial intelligence and innovative applications in education: the case of Turkey. *Journal of Information Systems and Management Research, 3*(1), 14-26.

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *Cambodian Journal of Educational Research*, *3*(1), 110-121. https://doi.org/10.2139/ssrn.4378735

Street, W., J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. Mckibben, T. Kanyere, A. Lentz, and Dunbar R.I. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint* arXiv: 2405:18870. https://arxiv.org/abs/2405.18870

Stribling, D., Xia, Y., Amer, M. K., Graim, K. S., Mulligan, C. J., & Renne, R. (2024). The model student: GPT-4 performance on graduate biomedical science exams. *Scientific Reports, 14*(1), 5670.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, *6*(1), 1-10. https://doi.org/10.37074/jalt.2023.6.1.17

Tapan Broutın, M. S. (2023). Matematik öğretmen adaylarının ChatGPT ile başlangıç deneyimlerinde sordukları soruların incelenmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 36*(2), 707-732. https://doi.org/10.19171/uefad.1299680

Trust, T., Whalen, J. & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education, 23*(1), 1-23.

Wen, J., & Wang, W. (2023). The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. *Clinical and Translational Medicine*, *13*(3), 1207. https://doi.org/10.1002/ctm2.1207

Wójcik, D., Adamiak, O., Czerepak, G., Tokarczuk, O., & Szalewski, L. (2024). A comparative analysis of the performance of ChatGPT4, Gemini and Claude for the Polish Medical Final Diploma Exam and Medical-

Dental Verification Exam. *MedRxiv.* https://doi.org/10.1101/2024.07.29.24311077

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

Author Information

Harun Bayer

https://orcid.org/0000-0002-8649-4026

Malatya Turgut Özal University

Boran, 44210 Battalgazi/Malatya

Turkiye

Contact e-mail: *harun.bayer@ozal.edu.tr*

Fazilet Gül İnce Aracı

https://orcid.org/0000-0001-5620-6911

Karamanoğlu Mehmetbey University

İbrahim Öktem Cd., 70100, Karaman

Turkiye

Gülşah Gürkan

https://orcid.org/0000-0003-0297-3060

Malatya Turgut Özal University

Boran, 44210 Battalgazi/Malatya

Turkiye