BMIJ
ISSN: 2148-2586

bmij (2024) 12 (4):989-1001
doi: https://doi.org/10.15295/bmij.v12i4.2445
Research Article

The potential of artificial intelligence models in tourism education: Exam performance and ethical discussions

Turizm eğitiminde yapay zekâ modellerinin potansiyeli: Sınav performansı ve etik tartışmalar

Levent Selman Göktaş[1]

[1] Assistant Prof. Dr., Harran University Faculty of Tourism, Department of Gastronomy and Culinary Arts, Sanliurfa, Türkiye,
leventselmangoktas@harran.edu.tr

ORCID: 0000-0001-6675-3759

Submitted: 1/10/2024

1st Revised: 10/12/2024

2nd Revised: 18/12/2024

3rd Revised: 21/12/2024

Accepted: 24/12/2024

Online Published: 25/12/2024

Abstract

This study aimed to compare the exam performances of ChatGPT Plus and Google Gemini Advanced in tourism management, tourism marketing, and tourism economics courses with the exam performance of undergraduate students. One hundred fifty students studying at Harran University Faculty of Tourism and completing their education in these three courses were selected and included in the exams with artificial intelligence models. In the exam, 25 questions were created for each course by academicians who are experts in their fields. The results show that ChatGPT has the highest overall accuracy rate and the lowest number of wrong answers. ChatGPT gave 21 correct answers in the tourism economics exam, Google Gemini 18, and students 16.6. In the tourism marketing exam, ChatGPT 19 and Google Gemini 18, students gave 14.9 correct answers. In the tourism management exam, ChatGPT answered 22 questions correctly, Google Gemini answered 14 questions, and students answered 16.3 questions correctly. In addition, the questions were categorised as short, long, easy, medium difficulty, complex questions, negative sentences, and scenario questions. When the results were analysed, ChatGPT was more successful in all categories. Although artificial intelligence language models are more effective than undergraduate students in certain exam conditions, this study underscores the need for further research to optimise and validate the use of these technologies in education. In addition, as a result of the research, it is thought that artificial intelligence language models can play a transformative role in tourism education in the future. An important finding has also emerged that ensuring the ethical and practical use of artificial intelligence technologies in academic settings requires responsible integration, human oversight, and more validation studies.

Keywords: ChatGPT, Google Gemini, Exam, Tourism

Jel Codes: L83

Öz

Bu çalışma ChatGPT Plus ve Google Gemini Advanced'ın turizm işletmeciliği, turizm pazarlaması ve turizm ekonomisi derslerindeki sınav performanslarını lisans öğrencilerinin sınav performansıyla karşılaştırmayı amaçlamıştır. Harran Üniversitesi Turizm Fakültesi'nde öğrenim görmekte olan ve bu üç ders özelinde eğitimini tamamlamış 150 öğrenci seçilerek yapay zeka modelleri ile birlikte sınavlara dahil edilmiştir. Sınavda her bir ders için alanında uzman akademisyenler tarafından 25 adet soru oluşturulmuştur. Sonuçlar, ChatGPT'nin en yüksek genel doğruluk oranına ve en düşük yanlış yanıt sayısına sahip olduğunu göstermiştir. ChatGPT turizm ekonomisi sınavında 21 doğru cevap verirken, Google Gemini 18, öğrenciler ise 16,6 doğru cevap vermiştir. Turizm pazarlaması sınavında ChatGPT 19, Google Gemini 18 öğrenciler ise 14,9 doğru cevap vermiştir. Turizm işletmeciliği sınavında ise ChatGPT 22, Google Gemini 14, öğrenciler ise 16,3 soruya doğru cevap vermiştir. Ayrıca sorular kısa, uzun, kolay soru, orta zorlukta soru, zor soru, olumsuz cümle ve senaryo soruları olarak kategorilendirilmiştir. Sonuçlar incelendiğinde tüm kategorilerde ChatGPT'nin daha başarılı olduğu görülmüştür. Yapay zeka dil modelleri belirli sınav koşullarında lisans öğrencilerinden daha etkili olsa da, bu çalışma bu teknolojilerin eğitimde kullanımını optimize etmek ve doğrulamak için daha fazla araştırmaya ihtiyaç olduğunu vurgulamaktadır. Ayrıca araştırma sonucunda yapay zeka dil modellerinin gelecekte turizm eğitiminde dönüştürücü bir rol oynayabileceği düşünülmektedir. Yapay zekâ teknolojilerinin akademik ortamlarda etik ve etkili kullanımını sağlamak için sorumlu entegrasyonun, insan gözetiminin ve daha fazla doğrulama çalışmasının gerektirdiği de önemli bir bulgu olarak karşımıza çıkmıştır.

Anahtar Kelimeler: ChatGPT, Google Gemini, Sınav, Turizm

JEL Kodları: L83

Citation: Göktaş, L.S., The potential of artificial intelligence models in tourism education: Exam performance and ethical discussions, bmij (2024) 12 (4): 989-1001, doi: https://doi.org/10.15295/bmij.v12i4.2445

Introduction

As technological developments in artificial intelligence have begun to be used in education, many studies have started to be conducted in this field (Popenici & Kerr, 2017; Chen, Chen, & Lin, 2020). In particular, GPT (Generative Pre-trained Transformer) models have played an important role in this development (Gimpel et al., 2023). GPT technology uses large amounts of publicly available digital content data to process and produce human-like handwritten texts. It also provides successful results in writing persuasive texts in many scientific fields (Grassini, 2023).

In recent years, many companies have focused on this technological development. However, OpenAI and Google Gemini have become famous by making significant advances in this field using chatbot technology (Timakov, 2023). The state-of-the-art ChatGPT OpenAI is a versatile tool that facilitates automated conversation and potentially makes human operators redundant (Kalla & Smith, 2023). In addition, ChatGPT has been used in education and training activities (Kasneci et al., 2023; Qadir, 2022), creating consistent content and articles (Castellanos-Gomez, 2023), preparing diet menus in gastronomy (Göktaş, 2023b), performing mathematical operations (Wardat, Tashtoush, AlAli & Jarrah, 2023; Frieder et al., 2023), language translation (Jiao, Wang, Huang, Wang & Tu, 2023), answering exam questions (Göktaş, 2023a) and programming code (Rahman & Watanobe, 2023). Gemini, developed by Google, generates real-time responses using natural language processing and machine learning. Google states that Gemini is successful in creative tasks, explaining complex topics and questions, and extracting information from various sources on the internet (Aydın, 2023). In addition, research has shown that Google Gemini performs well in many areas, such as answering exam questions (Vakilzadeh & Ghalejoogh, 2023), understanding and explaining visuals (Qin et al., 2023), diagnostic accuracy in the medical field (Hirosawa, Mizuta, Harada, & Shimizu 2023), and news verification (Caramancion, 2023).

While ChatGPT and Google Gemini succeeded in some of the exams in many scientific fields (Najafali et al., 2023; Koetsier, 2023), they failed in some exams with below-average results (Terwiesch, 2023; Ali et al., 2023; Angel, Patel, Alachkar & Baldi, 2023). Therefore, more research is needed to evaluate the performance of ChatGPT and Google Gemini in different types of exams (Ilgaz & Çelik, 2023). This study aims to fill this gap and to determine how ChatGPT and Google Gemini will perform in the exams taken with the student. In this study, ChatGPT and Google Gemini were chosen because they can capture complex language patterns and relationships, have a large number of parameters considered "large" ranging from hundreds of millions to hundreds of billions (Plevris, Papazafeiropoulos & Rios, 2023), and are significant language model-based bots with graphical user interfaces that are easy to use by a regular user (Urman & Makhortykh, 2023). Artificial intelligence language models such as ChatGPT and Google Gemini have potential advantages and risks for the future of education. One of these risks is that they can be used to cheat because they can produce personalised and real answers in online exams. As online exams become more common, ensuring their validity and reliability is important. For this reason, it is important to investigate the exam performance of artificial intelligence language models and compare their performance with real-time exams conducted with students. The fact that no comprehensive research has been conducted comparing the exam performance of students in tourism management, tourism marketing, and tourism economics courses offered at the Faculty of Tourism with the exam performance of artificial intelligence language models reveals the importance of the research in this respect. Accordingly, the study was conducted by determining the research questions.

In this context, the research questions of this study are as follows:

Question 1: Can ChatGPT Plus and Google Gemini Advanced be more successful than undergraduate students in the tourism management exam?

Question 2: Can ChatGPT Plus and Google Gemini Advanced be more successful than undergraduate students in the tourism marketing exam?

Question 3: Can ChatGPT Plus and Google Gemini Advanced be more successful than undergraduate students in the tourism economics exam?

Literature review

The impact of chatbots on learning, teaching, and assessment in higher education has been intensively discussed (Hien, Cuong, Nam, Nhung & Thang, 2018; Yang & Evans, 2019; Sandu & Gide, 2019; Essel, Vlachopoulos, Tachie-Menson, Johnson & Baah, 2022). Google Gemini and ChatGPT are two chatbots that can fulfil similar functions in this field (Ahmed et al., 2023). Considering the rising trend of both businesses (Google and OpenAI), it can be said that they are operating successfully (Urman & Makhortykh, 2023). As artificial intelligence technology develops, competition in this field becomes

natural and inevitable (Wang & Chen, 2018). However, among the competing companies, those who want to be successful will be the companies that can adapt quickly to new changes. Businesses have been in constant competition for years. However, the two competitors are strong and have entered a big competition. These competitors are Google Gemini and ChatGPT (Rahaman, Ahsan, Anjum, Rahman, & Rahman, 2023).

ChatGPT, a language model developed by OpenAI, has shown strong performance in various natural language processing tasks. It has been successful in tasks requiring translation, question answering, and instant reasoning (Brown et al., 2020). On the other hand, Google Gemini is another artificial intelligence tool introduced to humans as an artificial intelligence tool that can quickly respond to human questions by providing the highest data accuracy and avoiding erroneous inferences and wrong judgments based on assumption (Rahaman et al., 2023). These capabilities indicate that ChatGPT and Google Gemini have the potential to perform well in exams that include language-based questions and tasks (Ahmed et al., 2023). However, it should be noted that the performance of ChatGPT and Google Gemini in exams may vary depending on the specific context and requirements of the exam (Plevris et al., 2023; Ilgaz & Çelik, 2023). Although ChatGPT and Google Gemini have shown impressive learning capabilities during their development period, they still struggle in some areas (Plevris et al., 2023). ChatGPT and Google Gemini have achieved significant success in exams in different disciplines, such as law, health, and business (Metz & Collins, 2023; Skalidis et al., 2023; Patil, Huang, van der Pol & Larocque, 2023; Choi, Hickman, Monahan & Schwarcz 2023). These successes have also brought some precautions and prohibitions. For example, some universities and schools have banned ChatGPT (Rudolph, Tan & Tan, 2023; Yadava, 2023). The main reason for these bans is that students attempt to cheat in assignments, especially in online exams, and/or begin to atrophy in accessing and learning information using the artificial intelligence language model.

Artificial intelligence technology can potentially revolutionise teaching and learning methods in educational institutions (Kshirsagar et al., 2022; Eken, 2023). However, there are different opinions among educators and scientists about artificial intelligence tools. While some educators and scientists state that artificial intelligence tools such as ChatGPT and Google Gemini can make positive contributions to the future of education and research, some educators and scientists see it as a potential danger and state that it carries the risk of reducing educational activities and encouraging laziness among teachers and students due to the decrease in analytical skills (Grassini, 2023; Skavronskaya, Hadinejad & Cotterell 2023). When all these are considered in the context of education, it is essential to determine a responsible application strategy when using artificial intelligence tools such as ChatGPT or Google Gemini. Educators should consider the limitations and potential biases of artificial intelligence tools such as ChatGPT or Google Gemini and encourage their use as supplementary tools rather than replacements for human training (Halaweh, 2023). Strategies such as providing clear instructions to students, monitoring and verifying the accuracy of responses, and incorporating human feedback into the model's training can help reduce potential problems and increase the effectiveness of language modelling artificial intelligence tools in educational settings (Halaweh, 2023).

Advanced language models like ChatGPT and Google Gemini have the potential to negatively impact exams when not used responsibly and under appropriate supervision (Susnjak, 2022). While ChatGPT and Google Gemini can generate impressive text, several concerns and potential threats are associated with their use in the exam process. These concerns and potential threats are shown in Table 1.

Table 1: The Negative Impact of Artificial Intelligence (AI) Language Models on Exams

Concerns and potential threats	Description
The potential for unethical practices	The use of artificial intelligence language models in academia has raised concerns about the potential for unethical practices such as plagiarism and academic dishonesty (Lund et al., 2023; Sok & Heng, 2023; Malinka, Peresíni, Firc, Hujnák & Janus, 2023; Farrokhnia, Banihashem, Noroozi & Wals, 2023). Students may rely on these models to produce answers without fully understanding the content or engaging in critical thinking (Susnjak, 2022; Iskender, 2023; Yu, 2023).
Reliability and integrity	Artificial intelligence language models may produce technically correct answers but fail to address specific requirements or nuances of exam questions. This can lead to incomplete or irrelevant answers and inaccurate assessments of students' knowledge and understanding (Göktaş, 2023a; Kasneci et al., 2023; Sok & Heng, 2023).
Prejudice and justice	Uncontrolled use of artificial intelligence language models in exams can lead to biased or unfair results. These biases and unfair exam results can negatively impact specific student groups and undermine the assessment process's fairness (Susnjak, 2022; Currie, 2023; Cotton, Cotton & Shipway, 2023).
Lack of human feedback and guidance	Over-reliance on artificial intelligence language models in exams can reduce the role of tutors and feedback. Valuable instructor feedback and guidance can be reduced, hindering students' learning and development (Grassini, 2023).
Validity	Using artificial intelligence tools may raise concerns about the validity and reliability of exam results. Unresolved issues such as unreliability, low explainability, and bias in artificial intelligence may jeopardise the fairness of exam results and undermine the validity of national exams (Aloisi, 2023).
Impact on test preparation and teaching methods	Using artificial intelligence language models in scientific article writing can transform traditional exam preparation and teaching methods into a different dimension. This will positively and negatively affect the integrity of the education system (Farrokhnia et al., 2023).

When Table 1 is examined, one of the most important concerns posed by the artificial intelligence language model is the potential for plagiarism and academic dishonesty. Artificial intelligence language models can produce consistent and contextually relevant texts. This might make students compose answers or essays with the help of artificial intelligence language models without having a deeper understanding of the content or even thinking for themselves (Dergaa, Chamari, Zmijewski & Saad, 2023). Thus, exams intended to test a student's knowledge, comprehension, and analysis will not be able to serve their purpose. Another important problem is the inaccuracy and unreliability of the answer answered by the artificial intelligence language model. Artificial intelligence language models generate answers from extensive data from the internet, which can bring incorrect information (Dergaa et al., 2023). In an exam environment where accuracy and verification are critical, relying solely on artificial intelligence language models for answers can lead to students being given incorrect or misleading information (Göktaş, 2023a). Moreover, the artificial intelligence language model may not fully understand the information the exam questions require. While they may compose technically correct answers, they may fail to consider specific requirements or nuances of the question. This can lead to incomplete or irrelevant answers to questions, and therefore, students may be misled as to whether the information is correct (Göktaş, 2023a; Aloisi, 2023).

The exam performance of major language models such as ChatGPT and Google Gemini has recently become a research subject. These studies are important in understanding the strengths and weaknesses of AI language models and providing evidence and suggestions on how to ensure the validity and reliability of exams by introducing AI language models into our lives. Studies focusing on ChatGPT have shown that ChatGPT achieves good results in some science domains and average or poor results in others. A study by Gilson et al. (2022) found that ChatGPT performs at an average level on medical licensing exams. Furthermore, in one of the four exams evaluated, the results were found to be competitive with third-year medical students. Kung et al. (2023) found that ChatGPT achieves scores close to the cutoff required to pass the United States Medical Licensing Examination. ChatGPT performs poorly on exams and on exams where it performs well. ChatGPT performs below average in physics (Kortemeyer, 2023), medicine (Haverkamp, Tennenbaum & Strodthoff, 2023), and mathematics (Frieder et al., 2023). Similarly, Newton and Xiromeriti (2023) reported that ChatGPT performed below average on multiple-choice tests in several fields, including ophthalmology, law, economics, and physics.

The research on Google Gemini is minimal. In the studies conducted on Google Gemini, it was observed that while it achieved successful results in some areas, it achieved average or low results in others. Hirosawa et al. (2023) concluded that doctors generally succeed in case reports, but Google Gemini is not as successful as doctors. Nguyen et al. (2023) concluded that Google Gemini performed poorly in

the mathematics test. Phong et al. (2023) concluded that Google Gemini achieved an average result in the physics exam.

Methodology

This study is comparative research designed to evaluate the performance of ChatGPT Plus and Google Gemini Advanced in the exams taken with undergraduate students in tourism management, tourism marketing, and tourism economics courses. The reason for including these three courses in the research is that they are included in the curriculum of the departments of Harran University's tourism faculty, and the common opinion of 4 different academicians who are experts in their fields in selecting these courses is in this direction. Experimental design, one of the quantitative research methods, was used in the study. The study was conducted with Harran University Faculty of Tourism students in the 2024-2025 academic year. One hundred seventy-nine students have taken these courses at the Faculty of Tourism. Students who did not attend classes and education were excluded, and only students who took the courses were included in the study. In this context, the number of students who participated in the study was 150, and these students were students who had previously taken the courses "Tourism Management, Tourism Marketing, and Tourism Economics". Additionally, ChatGPT and Google Gemini models were included as exam participants. A special exam command was given to ChatGPT and Google Gemini to indicate that they were in the exam.

The data collection instrument utilised in this study consists of exam questions derived from the textbooks "Tourism Economics" (Bahar & Kozak, 2023), "Tourism Marketing" (Kozak, 2019), and "Fundamental Concepts and Practices in Tourism Management" (Akova, Kızılırmak & Tanrıverdi, 2015), which are associated with the courses in tourism economics, tourism marketing, and tourism management, respectively. These textbooks were selected because they are utilised as primary resources in the instruction of these courses. Each exam was prepared in alignment with the learning outcomes of the respective course. A question pool was created for each exam, and 25 multiple-choice exam questions were determined for each course based on the consensus of 4 academics who are experts in their fields. The exam questions created are new; students have not encountered them before. These exam questions have been determined to have a balanced distribution of easy-medium-difficulty. Since the students had taken these courses before, the exam date and topics were notified 2 months in advance, and they were asked to start preparing for the exam. The exams consisted of 3 sessions. The exams started in the specified exam hall and at the specified time. The duration of each exam was determined to be 30 minutes. The exams were held between October 8-10, 2024.

Table 2: Courses, Number of Questions, and Number of Students

Courses	Number of Questions	Number of Students
Tourism Management	25	150
Tourism Marketing	25	150
Tourism Economics	25	150

The study's implementation consisted of three stages: tourism management, marketing, and economics exams. Four academicians who are experts in their fields evaluated the exam results.

Table 3: Implementation Stages of the Research

Preparation Phase	Test questions were prepared and examined from source books determined by four academicians who are experts in their fields.
	ChatGPT and Google Gemini models are prepared to answer exam questions.
Data Collection Phase	Students took the exams on the designated dates.
	The same exam questions were asked separately to ChatGPT and Google Gemini.
	ChatGPT and Google Gemini's responses were recorded during the exam.
Evaluation Phase	Independent evaluators scored the student responses and the responses of the AI models.
	In the evaluation, the number of correct answers and the number of incorrect answers were taken into account.

Before the commencement of the research, ethical approval was obtained from the Social and Human Sciences Ethics Committee of Harran University, dated October 19, 2023, and numbered 2023/159. Student data were evaluated anonymously, adhering to principles of confidentiality.

Findings

The graphic below compares the accuracy and inaccuracy performances of ChatGPT and Google Gemini. As a result of the comparison, ChatGPT has a higher accuracy rate than Google Gemini (50), with 62 correct answers.

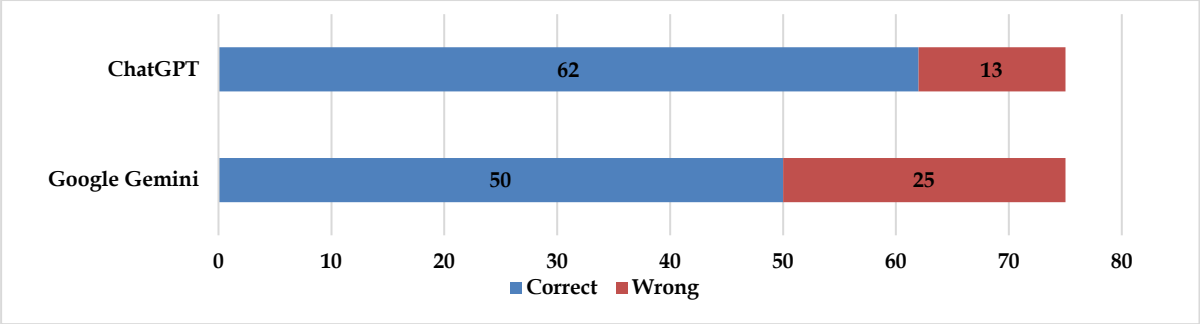

Figure 1: Number of Correct and Incorrect Answers of ChatGPT and Google Gemini in Exams

Regarding incorrect answers, ChatGPT made fewer mistakes than Google Gemini (25), with only 13 wrong answers. This data shows that ChatGPT was more successful than Google Gemini in answering the questions correctly. ChatGPT has a higher accuracy rate and a lower number of incorrect answers. ChatGPT is more effective than Google Gemini in certain exam conditions.

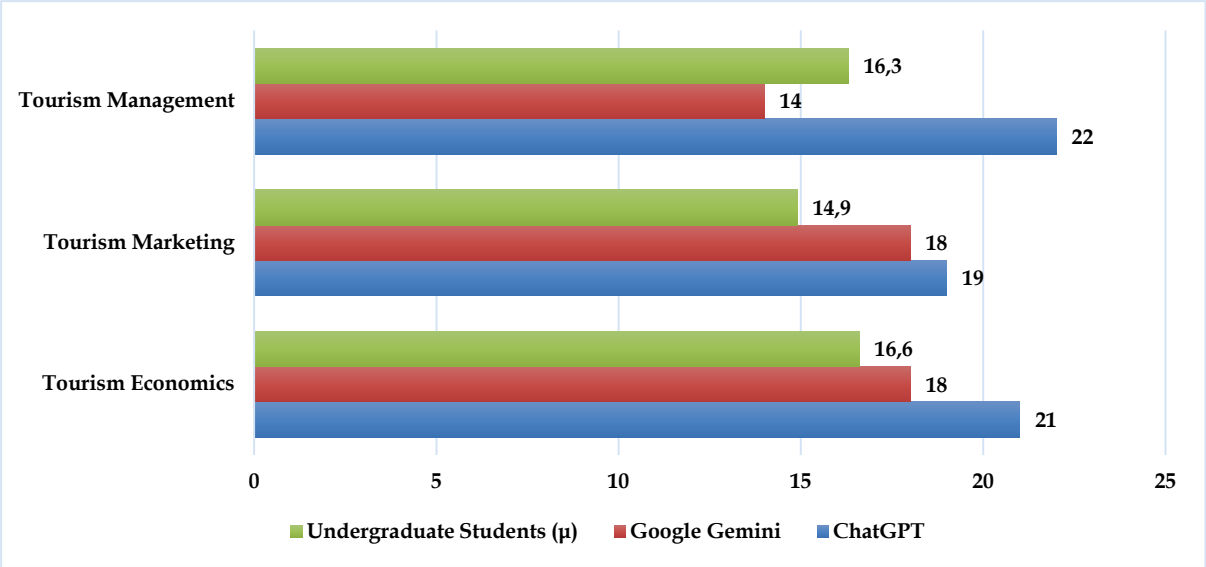

Figure 2: Total Number of Course-Based Correct on Exams for ChatGPT, Google Gemini, and Undergraduate Students (μ)

ChatGPT had the highest number of correct answers in the tourism economics exam and outperformed Google Gemini and undergraduate students. Google Gemini performed better than the students but not as well as ChatGPT. ChatGPT and Google Gemini performed similarly in the tourism marketing exam, but ChatGPT had a slightly higher number of correct answers. Both artificial intelligence models outperformed undergraduate students in the tourism marketing exam. ChatGPT had the highest number of correct answers in the tourism management exam. Google Gemini performed lower than undergraduate students in this category. ChatGPT's superior performance in this area was significantly better than the other two groups.

ChatGPT had the highest number of correct answers in all three categories and was the most successful model overall. Google Gemini outperformed undergraduate students in the tourism economics and marketing categories but underperformed them in the tourism management category. Undergraduate students generally had fewer correct answers compared to ChatGPT and Google Gemini, suggesting that artificial intelligence models can outperform students in specific academic exams.

Table 4: Overall Exam Performance of ChatGPT, Google Gemini, and Undergraduate Students (μ) According to Different Characteristics of the Questions

		ChatGPT		Google Gemini		Undergraduate Students (μ- mean)	
	Number of Questions	**Correct**	**Wrong**	**Correct**	**Wrong**	**Correct**	**Wrong**
Total Questions	**75**	62	13	50	25	47,8	27,2
Long questions (>50 words)	**29**	23	6	19	10	17,6	11,4
Short questions (≤50 words)	**46**	39	7	31	15	30,2	15,8
Easy question	**25**	23	2	20	5	19,5	5,5
Medium difficulty question	**25**	21	4	15	10	16,5	8,5
Difficult question	**25**	18	7	15	10	11,8	13,2
Negative sentence questions	**31**	27	4	23	8	24,1	6,9
Positive sentence questions	**44**	35	9	27	17	23,7	20,3
Scenario questions	**14**	10	4	7	7	6,3	7,7

Regarding overall performance, ChatGPT outperformed both Google Gemini and undergraduate students. ChatGPT has the highest accuracy rate and the least number of incorrect answers. ChatGPT again has the highest number of correct answers in the long questions category. Although Google Gemini outperformed undergraduate students, it did not perform as well as ChatGPT. This shows that ChatGPT understands complex and detailed questions better. On short questions, ChatGPT's performance was again at the top. While the number of correct answers was similar between Google Gemini and undergraduate students, ChatGPT was far more successful.

Another important criterion considered in the study was the difficulty levels of the questions. The difficulty levels of the questions were calculated using the item difficulty formula. The ratio of the number of people who answered an item correctly to the total number of people who took the test gives the item difficulty. Item difficulty indicates the features of the questions, such as easy, medium, and challenging questions. As this value approaches 1, the question becomes more manageable, and as it approaches 0, the question becomes more difficult; when it is close to 0.50, it is indicated that it is a question of medium difficulty (Badat, Usgu, Dinler, Bayramlar & Yakut, 2020). When the research results were examined according to the difficulty levels of the questions, it was understood that ChatGPT surpassed both Google Gemini and undergraduate students at all difficulty levels.

ChatGPT outperformed Google Gemini and undergraduates on negatively worded questions. Google Gemini achieved similar results with students, but ChatGPT had a higher accuracy rate. Regarding the number of correct answers, ChatGPT performed best on scenario questions. Google Gemini and undergraduates performed very similarly, with both groups underperforming ChatGPT. Similarly, ChatGPT was more successful in questions with positive expressions. Both Google Gemini and undergraduate students had lower correct answers than ChatGPT. This shows that ChatGPT can perceive questions with positive or negative sentence structures well and give correct answers.

The findings show that advanced AI language models can effectively understand and answer tourism education questions and have significant potential as learning and assessment tools. It is understood that AI language models have significant potential to improve tourism education by outperforming undergraduate students in tourism education exam scenarios, especially in answering complex and diverse questions. These findings show that AI has a significant transformative role in education.

Conclusion and suggestion

The results indicated that ChatGPT and Google Gemini can outperform undergraduate students in tourism exams. Furthermore, in all three categories, it turned out that ChatGPT's performance was even better, thus proving to be quite valuable for educational and training applications. These have provided critical data to assess the usage and potential of AI technologies in education. However, further research is required to verify these results and test their validity in various fields.

Among the categories, ChatGPT's performance had the highest number of correct responses and the least wrong. That would mean this result has proved that ChatGPT allows for correct comprehension and response to various questions compared to Google Gemini and undergraduate students. Whereas Google Gemini was proven to be better than undergraduate students in some categories, it was not as good as ChatGPT. These results also coincide with the research conducted by Cadiente, Chen, Kasselman, Pilkington (2024) and Cheong et al. (2024).

Where the artificial intelligence models did outperform the undergraduates was in long, complex, and negative sentences. The results show the potential to use artificial intelligence model technologies in educational and testing applications. On the other hand, such overuse of artificial intelligence models in exams may eventually deprive an instructor of her role. Exams are a place for expression of understanding on the part of a student and feedback on the part of an instructor. Dependence solely on artificial intelligence models for answering questions at exams inhibits instructors from giving meaningful knowledge and guidance; this may be considered one of the significant barriers to development that students face. For this reason, instructors must provide obvious guidelines, limitations, and prohibitions on adopting such tools as artificial intelligence models while taking exams. Besides that, instructor feedback and assessment should be hugely reflected in the process, avoiding possible biases.

Theoretical implications

When the research results were examined, it was seen that ChatGPT and Google Gemini had higher accuracy rates than the undergraduate students. This shows that AI language models successfully understand and answer academic questions. In addition, the artificial intelligence language model answered queries of long, complex, and negatively judged sentences better than undergraduate students. This shows that artificial intelligence language models are adapted for such questions. All these results helped us understand the performance of artificial intelligence language models on different question types. The fact that ChatGPT and Google Gemini were more successful than undergraduate students in terms of exam performance shows that artificial intelligence language models can surpass human performance in the future and set new standards in education. ChatGPT's superior performance on various question types (e.g., scenario-based) demonstrates that AI can support more profound understanding and problem-solving skills in tourism education.

This good exam performance of AI language models may also create some negative consequences. Students using AI language models to create answers and/or generate information during the exam may create results contrary to the learning process's nature. Using AI at all times may prevent students from developing essential competencies such as problem-solving, analytical thinking, and accessing information. In addition, the accuracy of the answers generated by AI language models is not guaranteed. This may cause students to fail. As a result, using AI language models in exams may negatively affect academic evaluation and exam justice unless ethical rules and a control mechanism are established.

Practical implications

ChatGPT and Google Gemini's superior performance shows that artificial intelligence language models can be used practically in educational institutions, such as course material development, answering exam questions, and student counselling. This can also increase efficiency in education. Artificial intelligence language models can help optimise students' learning processes by increasing correct exam answer rates. This can lead to a more accurate and reliable evaluation of exam results. The proliferation of artificial intelligence language models may change the role of educators in the future. Educators can provide students with more practical educational support by creating artificial intelligence-powered learning materials. The ability of AI to outsmart students in tourism-related subjects (economics, marketing, management) means it can help develop ancillary elements such as course materials and mock exams for challenging areas. Although they positively impact education, artificial intelligence language models are tools that require ethical and responsible usage strategies in exams and education. Educators must verify the accuracy of the answers provided by artificial intelligence and support students' learning processes. They must also guide students in using these tools correctly. The lack of clear rules and controls around the use of AI language models in exams can increase the potential for students to cheat and cheat, jeopardising the reliability of exam results. In order to manage this situation, faculty members should review exam formats and make students aware of the need to use this technology without exceeding ethical boundaries.

Since the study has some limitations, such as being conducted only at one university and not having regular exams for several consecutive years, future studies exploring the performance of artificial

intelligence models such as ChatGPT and Google Gemini across a broader range of disciplines and on more complex question types, as well as examining their effects on student learning outcomes, teaching practices, and academic honesty, could make significant contributions to the literature.

Peer-review:

Externally peer-reviewed

Conflict of interests:

The author has no conflict of interest to declare.

Grant Support:

The author declared that this study has received no financial support.

Ethics Committee Approval:

Ethics committee approval was received for this study from Harran University, Social and Human Sciences Ethics Committee on 09/01/2024 and 2024/48 document number.

References

Ahmed, I., Kajol, M., Hasan, U., Datta, P. P., Roy, A., & Reza, M. R. (2023). ChatGPT vs. Bard: A Comparative Study. *UMBC Student Collection*. https://doi.org/10.36227/techrxiv.23536290.v2

Akova, O., Kızılırmak, İ., & Tanrıverdi, H. (2015). *Turizm İşletmeciliği Temel Kavramlar ve Uygulamalar*. (9th Edition). Detay Yayıncılık. Ankara.

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., ... & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, 10-1227. https://doi.org/10.1227/neu.0000000000002551

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, 58(1), 98-110. https://doi.org/10.1111/ejed.12542

Angel, M., Patel, A., Alachkar, A., & Baldi, P. F. (2023). Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Pharmacy: A Comparative Study on the NAPLEX Exam. *bioRxiv*, 2023-06. https://doi.org/10.1101/2023.06.07.544055

Aydın, Ö. (2023). Google Bard Generated Literature Review: Metaverse. *Journal of AI*, 7 (1), 1-14. https://doi.org/10.61969/jai.1311271

Badat, T., Usgu, G., Dinler, E., Bayramlar, K., & Yakut, Y. (2020). Çoktan seçmeli sınavlarda kullanılan ölçme ve değerlendirme sisteminin uygulanması: Madde analiz örneği. *Hacettepe University Faculty of Health Sciences Journal*, *7*(3), 285-295. https://doi.org/10.21020/husbfd.629548

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural İnformation Processing Systems*, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Cadiente, A., Chen, J., Kasselman, L., & Pilkington, B. (2024). Large language models take on the AAMC situational judgment test: evaluating dilemma-based scenarios. *International Journal of Artificial Intelligence in Education.* https://doi.org/10.21203/rs.3.rs-4560463/v1

Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and Bard in news fact-checking. *arXiv preprint arXiv*:2306.17176. https://doi.org/10.48550/arXiv.2306.17176

Castellanos-Gomez, A. (2023). Good Practices for Scientific Article Writing with ChatGPT and Other Artificial Intelligence Language Models. *Nanomanufacturing*, 3(2), 135-138. https://doi.org/10.3390/nanomanufacturing3020009

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

Cheong, R. C. T., Pang, K. P., Unadkat, S., Mcneillis, V., Williamson, A., Joseph, J., ... & Paleri, V. (2024). Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2137-2143. https://doi.org/10.1007/s00405-023-08381-3

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT goes to law school. *Minnesota Legal Studies Research*. Paper No. 23-03, http://dx.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

Currie, G. M. (2023). Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?. In Seminars in Nuclear Medicine. *WB Saunders*. https://doi.org/10.1053/j.semnuclmed.2023.04.008

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, 40(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623

Eken, S. (2023). Ethic wars: Student and educator attitudes in the context of ChatGPT. *SSRN*. http://dx.doi.org/10.2139/ssrn.4365433

Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. International *Journal of Educational Technology in Higher Education*, 19(1), 1-19. https://doi.org/10.1186/s41239-022-00362-6

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1-15. https://doi.org/10.1080/14703297.2023.2195846

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *arXiv*. https://doi.org/10.48550/arXiv.2301.13867

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). How welldoes ChatGPT do when taking the medical licensing exams? The implications of large languagemodels for medical education and knowledge assessment. *medRxiv*. https://doi.org/10.1101/2022.12.23.22283901

Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., ... & Vandrik, S. (2023). Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers (No. 02-2023). *Hohenheim Discussion Papers in Business, Economics and Social Sciences*.

Göktaş, L. S. (2023a). ChatGPT Uzaktan Eğitim Sınavlarında Başarılı Olabilir Mi? Turizm Alanında Doğruluk ve Doğrulama Üzerine Bir Araştırma (Can ChatGPT Succeed in Distance Education Exams? A Research on Accuracy and Verification in Tourism). *Journal of Tourism & Gastronomy Studies*, 11(2), 892–905. https://doi.org/10.21325/jotags.2023.122

Göktaş, L. S. (2023b). The role of ChatGPT in vegetarian menus. *Tourism and Recreation*, 5(2), 79-86. https://doi.org/10.53601/tourismandrecreation.1343598

Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, 13(7), 692. MDPI AG. Retrieved from http://dx.doi.org/10.3390/educsci13070692

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, Volume 15, Issue 2, Article No: ep421 https://doi.org/10.30935/cedtech/13036

Haverkamp, W., Tennenbaum, J., & Strodthoff, N. (2023). ChatGPT fails the test of evidence-based medicine. *European Heart Journal-Digital Health*, https://doi.org/10.1093/ehjdh/ztad043

Hien, H. T., Cuong, P. N., Nam, L. N. H., Nhung, H. L. T. K., & Thang, L. D. (2018). *Intelligent assistants in higher-education environments: the FIT-EBot, a chatbot for administrative and learning support*. In

Proceedings of the 9th International Symposium on Information and Communication Technology (pp. 69-76). https://doi.org/10.1145/3287921.3287937

Hirosawa, T., Mizuta, K., Harada, Y., & Shimizu, T. (2023). Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. *The American Journal of Medicine*. https://doi.org/10.1016/j.amjmed.2023.08.003

Ilgaz, H. B., & Çelik, Z. (2023). The Significance of Artificial Intelligence Platforms in Anatomy Education: An Experience With ChatGPT and Google Bard. *Cureus*, 15(9). https://doi.org/10.7759/cureus.45301

Iskender, A. (2023). Holy or Unholy? Interview with Open AI's ChatGPT. *European Journal of Tourism Research*, 34, 3414. https://doi.org/10.54055/ejtr.v34i.3169

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv preprint arXiv*:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Kalla, D., & Smith, N. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology*, 8(3). Available at SSRN: https://ssrn.com/abstract=4402499

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

Koetsier, J. (2023). GPT-4 Beats 90% of Lawyers Trying to Pass the Bar. *Forbes.* (Access Date: 19.09.2023). https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/?sh=36a0d3053027

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19, 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kozak, M. & Bahar, O. (2023). *Turizm Ekonomisi*. (9th Edition). Detay Yayıncılık. Ankara.

Kozak, N. (2019). *Turizm Pazarlaması*. (8th Edition). Detay Yayıncılık. Ankara.

Kshirsagar, P. R., Jagannadham, D. B. V., Alqahtani, H., Noorulhasan Naveed, Q., Islam, S., Thangamani, M., & Dejene, M. (2022). Human intelligence analysis through perception of AI in teaching and learning. *Computational Intelligence and Neuroscience*, 2022. https://doi.org/10.1155/2022/9160727

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao,R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*. https://doi.org/10.1371/journal.pdig.0000198

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570-581. https://doi.org/10.1002/asi.24750

Malinka, K., Peresíni, M., Firc, A., Hujnák, O., & Janus, F. (2023). *On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree?.* In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1 (pp. 47-53). https://doi.org/10.1145/3587102.3588827

Metz C. & Collins K. (2023). El nuevo GPT-4: lo bueno y lo malo-The New York Times. Access date 20/09/2024 https://www.nytimes.com/es/2023/03/18/espanol/gpt-4-como-funciona.html

Najafali, D., Reiche, E., Araya, S., Camacho, J. M., Liu, F. C., Johnstone, T., ... & Fox, P. M. (2023). Bard Versus the 2022 American Society of Plastic Surgeons In-Service Examination: Performance on the Examination in its Intern Year. *In Aesthetic Surgery Journal Open Forum*. https://doi.org/10.1093/asjof/ojad066

Newton, P. M., & Xiromeriti, M. (2023). ChatGPT performance on MCQ-based exams. *EdArXiv*. https://doi.org/10.35542/osf.io/sytu3

Nguyen, P., Nguyen, P., Bruneau, P., Cao, L., Wang, J. & Truong, H. (2023). Evaluation of Mathematics Performance of Google Bard on The Mathematics Test of the Vietnamese National High School Graduation Examination. TechRxiv. *Preprint*. https://doi.org/10.36227/techrxiv.23691876.v1

Patil, N. S., Huang, R. S., van der Pol, C. B., & Larocque, N. (2023). Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Canadian Association of Radiologists Journal*, https://doi.org/10.1177/08465371231193716

Phong, N., Truong, H., Phuong, N., Philippe, B., Linh, C., & Jin, W. (2023). Google Bard's Physical Capabilities in Vietnamese High Schools. *SSRN*. http://dx.doi.org/10.2139/ssrn.4525863

Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *arXiv preprint*. ArXiv:2305.18618. https://doi.org/10.48550/arXiv.2305.18618

Popenici, S.A.D., Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12, 22 (2017). https://doi.org/10.1186/s41039-017-0062-8

Qadir, J. (2022). *Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education*. 2023 IEEE Global Engineering Education Conference (EDUCON). https://doi.org/10.1109/EDUCON54358.2023.10125121

Qin, H., Ji, G. P., Khan, S., Fan, D. P., Khan, F. S., & Gool, L. V. (2023). How good is Google Bard's visual understanding? An empirical study on open challenges. *Mach. Intell. Res*. 20, 605–613 (2023). https://doi.org/10.1007/s11633-023-1469-x

Rahaman, M. S., Ahsan, M. M., Anjum, N., Rahman, M. M., & Rahman, M. N. (2023). The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: An opinion article. *SRRN*. http://dx.doi.org/10.2139/ssrn.4351785

Rahman, M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. Applied Sciences, 13(9), 5783. *MDPI AG*. http://dx.doi.org/10.3390/app13095783

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, 6(1). https://doi.org/10.37074/jalt.2023.6.1.23

Sandu, N., & Gide, E. (2019). *Adoption of AI-Chatbots to enhance student learning experience in higher education in India*. In 2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1-5). IEEE. https://doi.org/10.1109/ITHET46829.2019.8937382

Skalidis, I., Cagnina, A., Luangphiphat, W., Mahendiran, T., Muller, O., Abbe, E., & Fournier, S. (2023). ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?. *European Heart Journal-Digital Health*, 4(3), 279-281. https://doi.org/10.1093/ehjdh/ztad029

Skavronskaya, L., Hadinejad, A., & Cotterell, D. (2023). Reversing the threat of artificial intelligence to opportunity: a discussion of ChatGPT in tourism education. *Journal of Teaching in Travel & Tourism*, 23(2), 253-258. https://doi.org/10.1080/15313220.2023.2196658

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *SSRN*. http://dx.doi.org/10.2139/ssrn.4378735

Susnjak, T. (2022). ChatGPT: The end of online exam integrity?. *arXiv preprint arXiv*:2212.09292. https://doi.org/10.48550/arXiv.2212.09292

Terwiesch, C. (2023). *Would chat GPT3 get a Wharton MBA. A prediction based on its performance in the operations management course*. Wharton: Mack Institute for Innovation Management/University of Pennsylvania/School Wharton. Available at: https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/

Timakov, K. A. (2023). *Comparison of current language models Google Bard and ChatGPT*. In. Modern strategies and digital transformations of sustainable development of society, education and science (pp. 168-171). https://doi.org/10.34755/IROK.2023.93.45.076

Urman, A. & Makhortykh, M. (2023). The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. *OSF Preprints*. https://doi.org/10.31219/osf.io/q9v8f

Vakilzadeh, A. & Pourahmad Ghalejoogh, S. (2023). Evaluating the Potential of Large Language Model AI as Project Management Assistants: A Comparative Simulation to Evaluate GPT-3.5, GPT-4, and Google-Bard Ability to pass the PMI's PMP test. *SSRN*. http://dx.doi.org/10.2139/ssrn.4568800

Wang, Y., & Chen, D. (2018). Rising sino-US competition in artificial intelligence. *China Quarterly of International Strategic Studies*, 4(02), 241-258. https://doi.org/10.1142/S2377740018500148

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

Yadava, O. P. (2023). ChatGPT-a foe or an ally? *Indian J. Thorac. Cardiovasc. Surg*. 39, 217–221. https://doi.org/10.1007/s12055-023-01507-6

Yang, S., & Evans, C. (2019). *Opportunities and challenges in using AI chatbots in higher education*. In Proceedings of the 2019 3rd International Conference on Education and E-Learning (pp. 79-83). https://doi.org/10.1145/3371647.3371659

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