EDUCATIONAL PROCESS
INTERNATIONAL JOURNAL

CORRESPONDENCE
Tijen Tülübaş

✉ *tijen.tulubas@dpu.edu.tr*

✉ *Faculty of Education, Kutahya Dumlupinar University, Evliya Celebi Campus, 43100, Kutahya, Türkiye.*

AUTHOR DETAILS
Additional information about the author is available at the end of the article.

RESEARCH ARTICLE

An Interview with ChatGPT on Emergency Remote Teaching: A Comparative Analysis Based on Human–AI Collaboration

Tijen Tülübaş ⓘ · **Murat Demirkol** ⓘ · **Tuncay Yavuz Ozdemir** ⓘ
Hakan Polat ⓘ · **Turgut Karakose** ⓘ · **Ramazan Yirci** ⓘ

ABSTRACT

Background/purpose – ChatGPT, a recent form of AI-based language model, have garnered interest among people from diverse backgrounds with its immersive capabilities. Using ChatGPT to support or generate scientific research has also created an ongoing debate over its advantages versus risks. The present study aimed to conduct an AI-enabled research process using ChatGPT so as to evaluate its potential to generate an accurate, clear, concise, and unbiased information as these are essential elements of rigorous scientific work.

Materials/methods – To achieve this aim, we worked on emergency remote teaching (ERT), which garnered significant interest due to its wide-spread use during the COVID-19 pandemic, and created opposing views, particularly in comparison to online teaching. We conducted a simultaneous query on ChatGPT-3.5 and 4 on five basic themes: (1) the definition and emergence of ERT, (2) the appropriateness of ERT for different grade levels, (3) a comparison between ERT and online teaching, (4) the possible outcomes of ERT, (5) the future prospects and uses of ERT, and we performed a comparative evaluation of these responses with regard to accuracy, clarity, conciseness, and potential bias. We also used Cohen's kappa to assess inter-rater agreement in our analysis.

Results – The results indicated that both versions were capable of generating accurate information without significant bias although the responses lacked depth and insight with being somewhat repetitive. As the level of judgment required by the query increased, the performance of ChatGPT-4 was much better; it provided clearer and more concise responses with a more synthesized and detailed categories of information on ERT.

Conclusion – Based on our results, we state that the cooperation of human and artificial intelligence is still warranted to ensure an accurate and reliable output from AI-based scientific queries. If ChatGPT is a plane with innovative technologies, there still needs to be a pilot in the cockpit to make use of these technologies in the best way so as to fly the plane safely to its destination.

Keywords – Artificial intelligence, ChatGPT, emergency remote teaching, OpenAI, generative AI, AI in education, chatbot.

To link to this article– https://dx.doi.org/10.22521/edupij.2023.122.6

1. INTRODUCTION

Since the release of ChatGPT (Chat Generative Pre-trained Transformer) by OpenAI Limited Partnership, San Francisco, USA on November 30, 2022, it has generated a lot of excitement and enthusiasm for individuals from diverse backgrounds about experimenting this new tool, and lively discussions have been taking place in many fields (Chen 2023). ChatGPT, as a recent form of generative artificial intelligence, is actually a large language model (LLM) with impressive capabilities in producing high-quality texts on a wide range of topics, generating coherent and human-like output when answering questions or statements, and thus allowing users to communicate with computers in a more natural and conversational way (Sabzalieva and Valentini, 2023).

Given the exponential interest in the capacity and capabilities of ChatGPT, it has also created an ongoing debate over the advantages versus the risks of advanced AI technologies in the scientific community. The discussions were two-fold. On one hand, it is considered that ChatGPT could aid academics when conducting research and writing papers with its diverse functionalities such as serving as a search engine that provides direct responses to queries, enabling users to bypass the need to manually sift through sources, generating preliminary drafts of written pieces, providing a useful ground for those struggling with writer's block, acting as an interlocutor in a brainstorming session to generate a range of novel ideas, and even promoting research equity and diversity through supporting non-English speaking researchers to better express and communicate their research ideas (Dowling and Lucey 2023; Gordijn and Have 2023; van Dis 2023). From this point of view, ChatGPT has been listed among the groundbreaking tools that can save time through aiding these crucial steps in research that require extensive efforts from human intelligence, and thus freeing up time for other critical research activities (Sallem, 2023).

On the other side of the debate, scholars have raised concerns regarding potential bias in ChatGPT's training datasets, which could limit its capabilities and result in factual inaccuracies, a phenomenon called artificial hallucination. In this case, ChatGPT could be generating a mix of true and fabricated academic texts, which raises accuracy and integrity concerns (Alkaissi and McFarlane 2023; Liebrenz et al. 2023; Sallam 2023). In addition to the possibility of creating inaccurate content, the risk of bias and discrimination, Borji (2023) listed several other potential issues associated with ChatGPT, including lack of transparency and reliability, cybersecurity concerns, ethical consequences, and societal implications. Similarly, due to the dearth of pertinent training data, ChatGPT is likely to produce unsatisfactory outcomes in tasks that demand logical or common-sense reasoning since it is unable to grasp the context and meaning of the text it is asked to generate (Lund and Wang 2023; Strubell et al. 2019).

Given that accuracy, clarity, coherence and bias avoidance are essential in scientific work, the accuracy and integrity of using ChatGPT for academic purposes still warrants more work since implications of ChatGPT on scientific writing is currently unknown (van Dis et al. 2023). Considering this void, the present study aims to conduct an AI-enabled research process using ChatGPT so as to evaluate its potential to generate an accurate, clear, concise, and unbiased information on a specified research topic. To achieve this aim, we preferred to work on emergency remote teaching (ERT), which has recently garnered significant interest in the scientific world due to its wide-spread use during the COVID-19 pandemic. ERT is also frequently contrasted with online teaching, and has generated opposing views in the relevant literature regarding the future prospects and uses of ERT in different stages of formal

education (Ferri et al. 2020). Thus, we consider that ERT offers a useful ground for an AI-based investigation trial that necessitates both theoretical and judgmental information.

Emergency remote teaching (ERT) is described as a temporary switch from face-to-face or blended instruction to fully remote teaching during crisis situations or emergencies with the purpose of providing a quick and reliable access to instruction rather than creating a comprehensive educational system (Hodges et al. 2020). The world has witnessed such a massive switch during the COVID-19 pandemic as the operations of almost all educational institutions around the world was suspended to avoid the health hazards of their citizens, and resorted to ERT to enable the continuity of instruction (Karakose 2021). As eloquently expressed by UNESCO (2020), education systems around the world faced an unprecedented challenge in the wake of massive school closures mandated as part of public health efforts to stop the spread of COVID-19, and as a result, education was delivered remotely through a mix of technologies in order to ensure continuity of curriculum-based study and learning for all.

Since the time it was utilized in response to the health threatening conditions of COVID-19 pandemic, ERT has garnered much debate among scholars with regard to its deficiencies as an unplanned form of instruction as well as its potential for supporting education in the case of any future crises, health threats, or natural disasters. As the lessons learned during the pandemic implied, ERT could cause serious psychological pressures and anxiety for not only students but for teachers, academics and families, necessitate alternative assessment and evaluation methods, amplify equity issues due to unequal access to technology and online tools, and also result in surveillance and data privacy concerns. Despite these hazards, scholars also contend that ERT should still be a matter of concern for educational specialists and scholars, and they should continue developing newer frameworks and assessing its potential implications for enhancing teaching/learning experience in such crises environments (Ferri et al. 2020; Hodges et al. 2020; Karakose 2021; Whittle et al. 2020).

2. METHODOLOGY

In the present study, we aimed to investigate how ChatGPT, as a recent AI-based innovation, responds to our search into emergency remote teaching (ERT) so as to evaluate the accuracy, clarity, and conciseness of information as well as the potential for any bias. We also wanted to conduct a comparative analysis of responses yielded by ChatGPT-3.5 and ChatGPT4 so that we can also identify any innovation brought by ChatGPT-4. With this purpose, we first conducted a comprehensive review of literature on ERT. Based on the literature on ERT (e. g., Bond et al. 2021; Hodges et al. 2020; Whittle et al. 2020), we developed a total of 23 open-ended questions to put to ChatGPT and retrieve fundamental information about ERT. We then conducted a panel meeting as the research team to discuss over this initial list of questions, and excluded questions that would potentially yield similar results. We also combined some of the questions to be able to reach more comprehensive answers. At the end of the panel discussion, all researchers agreed upon a total of 12 questions to be asked within 5 basic themes: (1) the definition and emergence of ERT, (2) the appropriateness of ERT for different grade levels, (3) a comparison between ERT and online teaching, (4) the possible outcomes of ERT, (5) the future prospects and uses of ERT.

Following this stage, we conducted a simultaneous search on ChatGPT-3.5 and ChatGPT-4 using the same set of questions in the same order. We used each question once, and did not pose any additional questions to limit our interference to this data collection process as researchers. As researchers, we first rated these responses for each basic category

independently for the four criteria mentioned earlier: (1) accuracy of information, (2) clarity of information, (3) conciseness of information, (4) the potential for bias. The scoring for each category was made using a trichotomous rating system: 1=completely inaccurate/unclear/unconcise/biased, 2=partly accurate/clear/concise/biased, 3=completely accurate/clear/concise/unbiased. The ratings were made for responses given by each version of ChatGPT. Following this stage, a focus group discussion was also held based on these same premises. Next, we conducted a focus group meeting where we compared and contrasted information provided for each basic research theme listed above, and evaluated the accuracy and breadth of information in comparison with published research on ERT as well as any indication of bias or stance on ERT as a method of teaching.

We also used Cohen's kappa to comparatively assess the accuracy, clarity, and conciseness of information, and the degree of bias. Cohen's kappa is a widely used statistic in social and medical sciences to quantitatively assess inter-rater agreement on a nominal scale, and the agreement between the raters is used as an indicator of the quality of the categories evaluated (Vieira et al. 2010; Warren 2015). Cohen's kappa, as a chance-corrected measure of inter-rater reliability, assumes two raters as alternate versions of a test, with their ratings being comparable to test scores and thus determines whether the degree of agreement between the two raters is greater than what could be expected by chance (Sun 2011). Therefore, it is commonly accepted as a robust measure of agreement compared to a straightforward computation of percentage (Ben-David 2008; Viera et al. 2010).

Cohen's kappa is calculated over the ratings of two assessors and has a value between 0 and 1. When both raters agree completely, Cohen's kappa takes the value of 1 whereas it takes a value of 0 when the level of agreement observed is the same as what would be expected by chance (Warrens 2015). Landis and Koch's (1977) benchmark guideline is often used to interpret values between 1 and 0. Accordingly, Cohen's kappa value between 0.00 and 0.20 indicates slight inter-rater agreement, between 0.21 and 0.40 fair agreement, 0.41and 0.60 moderate agreement, 0.61and 0.80 strong agreement, 0.81 and 1.00 almost perfect agreement. In the present study, the Cohen's kappa values were calculated over the assessments of two researchers (T.T. and T.K.), and the results were interpreted using Landis and Koch's (1977) benchmark.

3. RESULTS

In this section, we present the evaluations of responses given by ChatGPT-3.5 and ChatGPT-4 to queries made in the scope of five main themes: (1) the definition and emergence of ERT, (2) the appropriateness of ERT for different grade levels, (3) a comparison between ERT and online teaching, (4) the possible outcomes of ERT, (5) the future prospects and uses of ERT. Before elaborating on the assessment of results based on our trichotomous rating system for accuracy, clarity, conciseness and the possibility of bias, we preferred to include the responses by both versions of ChatGPT in Figures to enable transparency.

Our query on ChatGPT started with the definition and emergence of ERT, and the responses were as illustrated in Figure 1.

RESEARCH THEME: The Definition and Emergence of ERT

Model: GPT-3.5	Model: GPT-4
TI What is emergency remote teaching?	**TU** What is emergency remote teaching?
Emergency remote teaching is a form of online teaching that is implemented in response to a crisis or emergency situation, such as the COVID-19 pandemic. It involves using technology and online tools to deliver education remotely when it is not possible for students and teachers to meet in person. Emergency remote teaching is typically characterized by a rapid transition from traditional classroom teaching to online teaching, often with limited preparation and training. This mode of teaching may involve the use of video conferencing, learning management systems, online discussion forums, and other digital tools to facilitate learning and communication between teachers and students. While emergency remote teaching is not ideal, it can provide a way for students to continue their education during a crisis or emergency situation.	Emergency remote teaching (ERT) is a temporary shift in teaching methods, usually due to unexpected events or crises, that require educators to switch from traditional in-person instruction to remote or online teaching. This shift is typically implemented quickly, in response to situations like natural disasters, public health emergencies, or other disruptive events that prevent regular face-to-face instruction.

During emergency remote teaching, educators use various digital tools and platforms to deliver course content, facilitate communication, and provide feedback to students. ERT is distinct from well-planned online or distance learning, as it is often rapidly deployed without the benefit of extensive planning or preparation.

Key characteristics of emergency remote teaching include:

1. Quick transition: ERT is initiated in response to urgent circumstances, requiring educators to adapt their teaching methods on short notice.
2. Temporary nature: ERT is typically considered a short-term solution, with the expectation that traditional in-person instruction will resume once the crisis or disruption has passed.
3. Focus on essential content: Due to the constraints and challenges of ERT, educators often prioritize the most critical course content and learning outcomes to ensure students' progression in their studies. |
| **TI** When did emergency remote teaching first start? | **TU** When did emergency remote teaching first start? |
| Emergency remote teaching has been used in response to various crises and emergencies over the years, such as natural disasters, conflicts, and public health emergencies. However, the term "emergency remote teaching" became widely used during the COVID-19 pandemic in 2020, when schools and universities around the world were forced to close their physical campuses and transition to online learning almost overnight. This sudden shift to remote learning was necessary to protect the health and safety of students and staff, and to help prevent the spread of the virus. While remote learning has been used in various forms for years, the scale and scope of emergency remote teaching during the COVID-19 pandemic was unprecedented. | Emergency remote teaching (ERT) does not have a specific starting point, as it is a response to various crises or unexpected events that require a shift from traditional in-person instruction to remote or online teaching. Throughout history, there have been instances where educators had to adapt their teaching methods due to disruptions caused by natural disasters, wars, public health emergencies, or other events.

However, the widespread use of ERT became more prominent with the advent of digital technologies and the internet, which allowed for more seamless transitions to remote teaching. A notable recent example of ERT on a global scale occurred during the COVID-19 pandemic, which began in late 2019 and continued into 2020 and 2021. During this period, schools and educational institutions around the world had to quickly shift to remote teaching as a means of maintaining educational continuity while adhering to public health guidelines and restrictions.

In summary, ERT has been employed in various forms and degrees throughout history in response to crises and disruptions. The COVID-19 pandemic, however, marked a particularly significant and widespread implementation of ERT, facilitated by advancements in digital technologies and online learning platforms. |
| **TI** When did emergency remote teaching emerge as a method of teaching? | **TU** When did emergency remote teaching emerge as a method of teaching? |
| Emergency remote teaching emerged as a method of teaching in response to crises and emergencies that made it difficult or impossible for students and teachers to meet in person. For example, after Hurricane Katrina devastated New Orleans in 2005, emergency remote teaching was used to help displaced students continue their education while schools were closed or damaged. Similarly, during the Ebola outbreak in West Africa in 2014-2015, emergency remote teaching was used to provide education to students who were unable to attend school due to the risk of infection. However, it was the COVID-19 pandemic in 2020 that brought emergency remote teaching into widespread use on a global scale, as schools and universities around the world were forced to close their physical campuses and transition to online learning almost overnight. The sudden and unprecedented nature of the pandemic led to emergency remote teaching becoming a necessary and widely adopted method of teaching almost overnight. | Emergency remote teaching (ERT) emerged as a method of teaching in response to various crises or unexpected events that necessitated a shift from traditional in-person instruction to remote or online teaching. It is not a well-defined or planned teaching methodology but rather an improvised response to challenging circumstances.

While instances of ERT have occurred throughout history, it gained prominence with the advent of digital technologies and the internet. This allowed for easier transitions to remote teaching during disruptive events, as educators could leverage digital tools and platforms to maintain continuity in education.

The most significant and widespread implementation of ERT in recent history occurred during the COVID-19 pandemic, which began in late 2019 and continued into 2020 and 2021. In response to the pandemic, schools and educational institutions worldwide had to quickly shift to remote teaching to ensure the safety of students and educators while adhering to public health guidelines and restrictions.

In summary, ERT emerged as an improvised teaching method in response to crises and disruptions, and it has evolved alongside the development of digital technologies and online learning platforms. The COVID-19 pandemic marked a critical point in the use of ERT on a global scale. |
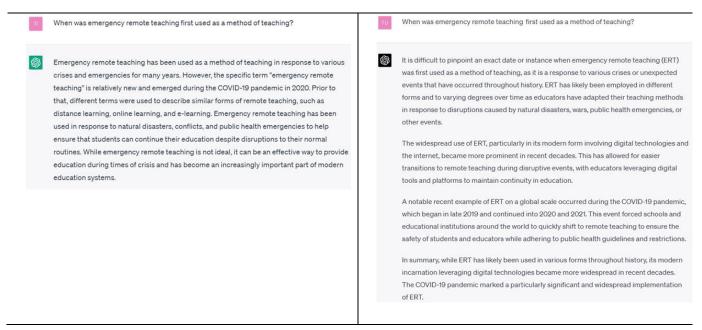
Figure 1. Sample excerpts from responses of ChatGPT-3.5 & ChatGPT-4 for the definition and emergence of ERT (generated on April, 13, 2023)

The researchers rated the responses by both ChatGPT-3.5 and ChatGPT-4 based on the trichotomous rating system. The assessment of accuracy for ChatGPT-3.5 revealed a mean value of 2,8 out of 3, indicating a high level of accuracy. Similarly, the mean value for clarity and conciseness were 2,4 and 2,8 respectively. With regard to the possibility of bias, researchers classified the responses as being almost unbiased with a mean value of 2,4. The results of Cohen's kappa showed that the inter-rater agreement on the nominal scale was almost perfect, indicated with a kappa value of 0.86.

When the responses by ChatGPT-4 with regard to the definition and emergence of ERT were evaluated, the mean values of researcher's assessment were 2,8 for accuracy and clarity, 2,6 for conciseness and the possibility of bias. The kappa value of 0,91 indicated almost perfect inter-rater agreement for these responses by ChatGPT-4.

A closer scrutiny into these responses shows that ChatGPT-3.5 has a clearer focus that ERT is a form of online teaching while ChatGPT-4 tends to define it as a form of remote teaching mostly depending on the use of digital technologies. Similarly, although both versions form links between the emergence of ERT and COVID-19 outbreak, ChatGPT-4 tends to avoid forming direct relationships between ERT and COVID-19. This is particularly evident in its definition of ERT, where it makes no reference to COVID-19 while ChatGPT-3.5 does. Likewise, ChatGPT-4 gives a much thorough and comprehensive definition of ERT with a more to-the-point choice of terminology used in the educational literature. Another interesting result is that ChatGPT prefers to use an acronym for emergency remote teaching, that's ERT, although the researchers avoided using acronyms during the query, which could be interpreted as ChatGPT-4 having stronger capacity for making judgments in presenting information to the interrogator.

Our next query on ChatGPT addressed the appropriateness of ERT for different grade levels, and the responses were as illustrated in Figure 2.

RESEARCH THEME The Appropriateness of ERT for Different Grade Levels

Model: GPT-3.5	Model: GPT-4

Figure 2. Sample excerpts from responses of ChatGPT-3.5 & ChatGPT-4 for the appropriateness of ERT for different grade levels (generated on April, 13, 2023)

The assessment of researchers for the accuracy of responses provided by both ChatGPT-3.5 revealed a mean value of 3, indicating full accuracy. Similarly, the mean value for clarity and conciseness of responses by ChatGPT-3.5 were 2,6 while it was 2,8 for the responses by ChatGPT-4. With regard to the possibility of bias, researchers classified the responses as being almost unbiased with a mean value of 2,8. The results of Cohen's kappa showed that the inter-rater agreement on the nominal scale was almost perfect, indicated with a kappa value of 0.94.

Although ChatGPT-3.5 and ChatGPT-4 gave similar responses with regard to the appropriateness of ERT for different grade levels, ChatGPT-4 seems to offer a more synthesized version of the same information, which makes its responses more intelligible and fluent.

In the third stage of query, we asked ChatGPT about the similarities and differences between ERT and online teaching, and the responses were as illustrated in Figure 3.

RESEARCH THEME: The Comparison between ERT and Online Teaching

Model: GPT-3.5	Model: GPT-4

Figure 3. Sample excerpts from responses of ChatGPT-3.5 & ChatGPT-4 for the comparison between ERT and online teaching (generated on April, 13, 2023)

Our query of similarities and differences between ERT and online teaching certainly required a more detailed analysis and synthesis, and as illustrated by the responses in Figure 3, ChatGPT-4 demonstrated a better performance in this regard. In this domain, the assessment of accuracy for ChatGPT-3.5 revealed a mean value of 2,4 while the result for clarity was 2,2 and 2 for conciseness. With regard to the possibility of bias, researchers classified the responses as being almost unbiased with a mean value of 2,4. The results of Cohen's kappa showed that the inter-rater agreement on the nominal scale was almost perfect, indicated with a kappa value of 0.89.

When the responses by ChatGPT-4 were evaluated, the mean values of researcher's assessment were 3 for accuracy, 2,8 for clarity and conciseness, and 2,6 for the possibility of bias. The kappa value of 0,92 indicated almost perfect inter-rater agreement for these responses by ChatGPT-4.

A closer scrutiny into these responses shows that ChatGPT-3.5 tends to give more general responses and often lacks details that could make the information clearer and vivid in the interrogator's mind. For instance, it mentions that ERT requires different teaching methods but lacks any examples of these methods. However, ChatGPT-4 provides more detailed information about the different pedagogies and techniques required in the practice of ERT. Similarly, ChatGPT has better capacity to categorize information and present it in a clearer and concise fashion. In addition, ChatGPT-3.5 focuses on assessment in both similarity and difference domain while ChatGPT-4 prefers to focus more on pedagogic differences, and also underlines the fact that providing technical or professional support during ERT could be

more difficult as compared to online teaching, which is not pointed by ChatGPT-3.5. These examples imply that ChatGPT-3.5 and ChatGPT-4 tend to have a different foci in this comparative analysis.

In the fourth stage, we posed questions about the positive and negative outcomes of ERT, and the responses by the two versions of ChatGPT were as illustrated in Figure 4.

RESEARCH THEME: The Outcomes of ERT

Model: GPT-3.5	Model: GPT-4

Figure 4. Sample excerpts from responses of ChatGPT-3.5 & ChatGPT-4 for the outcomes of ERT (generated on April, 13, 2023)

As our query went deeper on ERT, it seemed to us that both versions of ChatGPT were likely to produce more comprehensive and detailed responses. The accuracy, clarity and conciseness of responses were categorized by researchers as being almost perfect with a mean value of 2,6 for ChatGPT-3.5, and 3 for ChatGPT-4. As our fourth query addressed the possible positive and negative outcomes of ERT, it required stronger judgment on the use of ERT. Therefore, ChatGPT-4 employed a slightly better performance than ChatGPT-4. However,

both versions attempted to avoid providing biased information, as indicated by their responses including both the positive and negative aspects of resorting to ERT as a method of teaching. The mean value of 2,8 out of 3 for the possibility of bias showed that researchers evaluated ChatGPT-4 responses for this domain as being almost unbiased. The results of Cohen's kappa also showed that the inter-rater agreement on the nominal scale was almost perfect, indicated with a kappa value of 0.91.

When the positive and negative outcomes of ERT as listed by ChatGPT-3.5 and ChatGPT-4 reviewed, it can be clearly seen that they actually say the same thing using different words or phrases. However, ChatGPT evidently generates more detailed responses including outcomes that are not mentioned by ChatGPT-3.5. For instance, for the positive outcomes of ERT, ChatGPT-4 lists improvement of health and safety, digital literacy of teachers and students and the inclusivity in education, but ChatGPT-3.5 does not. Similarly, in the negative outcomes domain, Chat GPT 4 mentions increased stress and workload of teachers in addition to other educational outcomes. On the other hand, ChatGPT continues to focus on the likelihood of decreased quality of assessment during ERT.

In the final stage of our query, we interrogated ChatGPT about the future prospects and uses of ERT, and the responses were as illustrated in Figure 5.

RESEARCH THEME: The Future Prospects and Uses of ERT

Model: GPT-3.5	Model: GPT-4

Figure 5. Sample excerpts from responses of ChatGPT-3.5 & ChatGPT-4 for the future prospects and uses of ERT (generated on April, 13, 2023)

With regard to the future prospects and uses of ERT, responses by the two versions of ChatGPT were quite different although they shared some common grounds as well. Although the mean values of researcher assessment for the accuracy of information was high (2,8 for ChatGPT-3.5 and 3 for ChatGPT-4), ChatGPT-4 provided more concise and thorough information. The mean values for the clarity and conciseness of ChatGPT-4 responses were 2,8 while it was 2 for ChatGPT-3.5 responses. The Cohen's kappa value of 0.86 also indicated a high level of inter-rater agreement with this regard.

Articulating responses for the possible implications of ERT in the future clearly demands more judgmental stance based on existing knowledge, and ChatGPT-4 seems to have gone way forward compared to its previous version. Although some aspects of their responses are similar, ChatGPT underlines the potential of ERT to bring some solutions to the ever-lasting problems of education such as social justice, customized teaching, resilient education systems in the face of unexpectedly and fast changing world. In fact, what makes ChatGPT-4 responses

look more convincing seems to be its capacity to offer a more comprehensive and concise categories of information.

4. DISCUSSION

The current study aimed to provide an account of the responses and perspectives of the two versions of ChatGPT (3.5 and 4) with respect to emergency remote teaching, taking into account scientific accuracy, clarity, conciseness, and potential biases as the essential components of a scientific work. The present investigation sought to conduct an initial assessment of a novel technological advancement grounded in artificial intelligence, namely ChatGPT, which could potentially serve as a transformative agent for the acquisition and dissemination of knowledge in the foreseeable future.

Our descriptive assessment of the responses by ChatGPT-3.5 and 4 indicated that both versions were capable of generating accurate information without significant bias when the topic under investigation, ERT in the present case, is clearly discussed in the literature without significant controversy. However, it is also noteworthy that these responses are not only based on published research but also on a huge variety of data provided on the internet (Gilat and Cole 2023). Given that scientific knowledge grows on previous evidence and thus references to previous publications is essential, these automatic, AI based responses cannot be directly used in scientific writing as ChatGPT does not currently provide correct citing resources, and can even fabricate non-existent references (Chen 2023).

Considering a previous finding that ChatGPT performed greatly in generating queries for systematic reviews with high precision, despite cautions for the scope and transparency of analysis (Aydın and Karaaslan 2022; Wang et al. 2023), the accuracy, clarity and conciseness of information yielded by our search should not be surprising as most of the questions we put to ChatGPT could be answered after a thorough review of the literature. However, as the level of judgment required by the query increases, the performance of ChatGPT-4 seemed to be well ahead of ChatGPT-3.5's, particularly with regard to conciseness and clarity of answers. ChatGPT-4 provided a more synthesized and detailed categories of information with a larger scope of view on ERT, and with a more technical language. Yet, the responses produced by both versions still lacked depth and insight with mostly being repetitive as previously underlined (Cahan and Treutlein 2023; Gordijn and Have 2023; Liebrenz et al. 2023). From this perspective, our findings also conflict with the previous assumption that ChatGPT could generate over-detailed or excessive content that would pose additional burden on the researchers (Aczel and Wagenmakers 2023; Chen 2023; Stokel-Walker and van Noorden 2023), but supports the view that using proper prompts and posing clear questions matter in carefully supervising ChatGPT to produce to-the-point responses (Sallam 2023; Wang et al. 2023).

With regard to our query on the definition and emergence of ERT, both versions of ChatGPT generated convincing responses. In a frequently cited article, Hodges et al. (2020, p. 6) defines ERT as "a temporary shift of instructional delivery to an alternate delivery mode due to crisis circumstances. It involves the use of fully remote teaching solutions for instruction or education that would otherwise be delivered face-to-face or as blended or hybrid courses and that will return to that format once the crisis or emergency has abated". ChatGPT generated very similar versions of this definition. With respect to the emergence of ERT, though, ChatGPT-3.5 gave different examples of times it was used a s a method of teaching such as the Ebola outbreak in Africa during 2014-2015. Indeed, several studies were published on the use of television and radio broadcasting as a means of delivering instructions

(e.g. Darvas 2017; Hallgarten 2020; Smith 2021). Interestingly, ChatGPT-4 avoided these examples, and only mentioned COVID-19 pandemic as a typical example of a time where ERT was widely employed. This focus on COVID-19 was also evident in the responses of ChatGPT-3.5.

The differences between online teaching and ERT was another topic of hot debates in the scientific world following the COVID-19 outbreak (Barbour et al. 2020; Bozkurt and Sharma 2020; Ferri et al. 2020; Karakose 2021b). Therefore, our second query was about the similarities and differences between ERT and online teaching, where both versions of ChatGPT underlined that ERT lacked many of the essential elements of online teaching such as social contact and student interaction, carefully-prepared, high-quality digital materials, improved digital skills for both teachers and students, easy and equitable access to digital devices and digital content (Adam et al. 2017; Delen and Liew 2016). These results were really promising, particularly regarding that the answers to this query required comparisons and deeper synthesis of information.

The third step in our query on ChatGPT-3.5 and 4 were about the positive and negative outcomes of ERT. It was noteworthy that both versions approached the two questions, about the positive and the negative outcomes, from both angles. In other words, when we asked about the positive outcomes, they both mentioned that negatives outcomes should also be regarded, and vice versa. We interpreted this as an attempt to display an objective and unbiased stance during their interaction with the interlocutor. This could be assumed to be good sign of potential capability of generating more objective and accurate information in the forthcoming versions of AI-generated LLMs.

In the final set of our query, we questioned ChatGPT about the potential uses and benefits of ERT in the future. As can be expected, the first responses of both versions addressed its use in possible crises situations such as natural disasters or other health-threatening outbreaks. However, when we asked for other possible benefits of ERT without experiencing such emergencies, both versions were able to articulate judgments regarding the possible uses of ERT in solving some of the long-lasting problems of education such as acquiring equitable standards for learners with special needs, providing customized learning appropriate for the unique needs of students, and establishing a more flexible learning environment in the contemporary fast-changing context of schools. Although these responses seemed a bit over-enthusiastic in the face of existing research evidence, they provide a useful ground for developing different perspectives into the improvement of ERT experience in future instances of its use. Another point we should make is that, in this domain of our query, ChatGPT-4 generated more comprehensive and detailed responses as compared to ChatGPT-3.5. This could be due to the fact that ChatGPT-4 demonstrates better performance in areas such as reasoning, following user intent, knowledge retention and synthesis compared to earlier models such as ChatGPT-3.5 and the responses generated by ChatGPT-4 were preferred those generated by ChatGPT-3.5 over 70% of the trials (OpenAI 2023).

5. CONCLUSION

Since the release of ChatGPT, particularly its most recent version GPT-4, several potential use cases have been postulated or tested such as designing webpages, writing minutes, creating catalogs, writing newspaper articles, songs, poems, jokes, and scripts, providing therapy, organizing unstructured data, generating queries, creating automated applications, and answering open-ended analytical questions (Taecharungroj 2023). While the improvement in the AI technologies have already started to make many of these tasks

unexclusive, whether these developments could make some professionals redundant has become another object of curiosity (Greene 2022; Warner 2023). In fact, innovative technologies like ChatGPT will certainly have both favorable and unfavorable consequences. In the realm of science, the technology that underlies ChatGPT could fundamentally transform how people search for information by delivering immediate and in-depth results, and ChatGPT has already earned reputation as a groundbreaking tool that can serve as a research assistant, idea generator, or language editor (Sabzalieva & Valentini, 2023). Yet, as underlined by several scholars, more investigation is necessary to assess the content of ChatGPT to comprehend its potential impact on the advancement of scientific work (Checcucci et al. 2023; Gordijn and Have 2023; Sallam 2023). The results of the present study support this assumption despite showing that ChatGPT has a strong potential to aid research through generating accurate, clear, and concise information mostly free of bias. However, as cautioned earlier, ChatGPT is not governed by ethical principles and does not take the responsibility of content it produces, and as the outcomes it produces comes from the statistical analysis of billions of texts on the internet, it is still likely to reflect any cognitive bias found in this information (Sabzalieva and Valentini 2023).

The originality of the current study lies in the fact that it was one of the first studies to conduct a comparative evaluation of content produced by ChatGPT-3.5 and 4, and the results were actually very impressive. Our findings showed that the recent version of ChatGPT seems to offer better guidance for scientific queries, complying with Dowling and Lucey's (2023) expectation that ChatGPT-4 'promises a truly revolutionary language model due to advances in algorithms and over 600 times greater testing parameters' (p. 5). Yet, in its current state, the utility of ChatGPT for academic purposes seems to be limited, but we should also acknowledge that it is certainly a groundbreaking step in the development of tools to enhance scientific work (Gordijn and Have 2023). Therefore, we want to reiterate Sallam's (2023) and Kasneci et al.'s, (2023) claim that the cooperation of human and artificial intelligence is still warranted to ensure an accurate and reliable output from AI-based queries. To explain with an allegory, if we consider ChatGPT as a plane with innovative technologies, there still needs to be a pilot in the cockpit to make use of these technologies in the best way so as to fly the plane safely to its destination.

DECLARATIONS

REFERENCES

Aczel, B., and Wagenmakers, E. 2023. Transparency guidance for ChatGPT usage in scientific writing. *PsyArXiv*, Preprint.

Adam N.L., Alzahri F.B., Cik S.S., Abu Bakar N., and Mohamad Kamal N.A. 2017. Self-regulated Learning and Online Learning: A Systematic Review. In *Advances in Visual Informatics*. *Lecture Notes in Computer Science*. Edited by H. B. Zaman, P. Robinson, A. F. Smeaton, T. K. Shih, S. Velastin, T. Terutoshi, A. J. Nazlena, and M. Ali. New York, NY: Springer, vol. 10645, pp.143-154.

Alkaissi, H., and McFarlane, S. I. 2023. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 15(2). https://doi.org/10.7759/cureus.35179

Aydın, Ö., and Karaarslan, E. 2022. OpenAI ChatGPT generated literature review: digital twin in healthcare. In *Emerging Computer Technologies 2*. Edited by Ö. Aydın. İzmir: İzmir Akademi Dernegi, pp. 22-31. http://dx.doi.org/10.2139/ssrn.4308687

Barbour, M. K., LaBonte, R., Hodges, C.B., Moore, S., Lockee, B.B., Trust, T. Bond, M.A., Hill, P., and Kelly, K. (2020). Understanding pandemic pedagogy: Differences between emergency remote, remote, and online teaching. *Canadian eLearning Network*. https://vtechworks.lib.vt.edu/handle/10919/101905

Ben-David, A. 2008. Comparison of classification accuracy using cohen's weighted kappa. *Expert Systems with Applications* 34(2): 825-832.

Borji, A. A. 2023. Categorical archive of ChatGPT failures. *arXiv*:2302.03494.

Bozkurt, A., and Sharma, R. C. 2020. Emergency remote teaching in a time of global crisis due to coronavirus pandemic. *Asian Journal of Distance Education* 15(1): i-vi.

Cahan, P., and Treutlein, B. 2023. A Conversation with ChatGPT on the role of computational systems biology in stem cell research. *Stem Cell Reports* 18(1): 1-2.

Checcucci, E., Verri, P., Amparore, D., Cacciamani, G.E., Fiori, C., Breda, A., Porpiglia, F. 2023. Generative pre-training transformer chat (chatgpt) in the scientific community: The train has left the station. *Minerva Urology and Nephrology* 75(2): 131-133. https://doi.org/10.23736/S2724-6051.23.05326-0

Chen, T. J. 2023. ChatGPT and other artificial intelligence applications speed up scientific writing. *Journal of the Chinese Medical Association* 86: 351-353. https://doi.org/10.1097/JCMA.0000000000000900.

Darvas, P. 2017. *Liberia - Education Sector - Ebola Recovery and Reconstruction Project: P154880 - Implementation Status Results Report : Sequence 02 (English)* http://documents.worldbank.org/curated/en/171961488217795401/pdf/ISRisclosableP 154880-02-27-2017-1488217785586.pdf

Delen, E., and Liew, J. 2016. The use of interactive environments to promote self-regulation in online learning: A literature review. *European Journal of Contemporary Education* 15(1): 24-33. https://eric.ed.gov/?id=EJ1095976

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The bananarama conjecture. *Finance Research Letters* 53: 103662. https://doi.org/10.1016/j.frl.2023.103662

Ferri, F., Grifoni, P., and Guzzo, T. 2020. Online learning and emergency remote teaching: opportunities and challenges in emergency situations. *Societies* 10(4): 86-95. https://doi.org/10.3390/soc10040086

Gilat, R., and Cole, B. J. 2023. How will artificial intelligence affect scientific writing, reviewing and editing? The Future is Here…. *Arthroscopy* 39(5): 1119-1120.

Gordijn, B., and Have, H. T. (2023). ChatGPT: Evolution or revolution?. Medicine, Health Care and Philosophy, 1-2.https://doi.org/10.1007/s11019-023-10136-0

Greene, J. 2022. Will ChatGPT make lawyers obsolete? (Hint: Be afraid). *Reuters* (Dec 09, 2022).

Hallgarten, J. 2020. *Evidence on efforts to mitigate the negative educational impact of past disease outbreaks K4D Helpdesk Report 793*. Reading, UK: Education Development Trust. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/15202

Hodges, C. B., Moore, S., Lockee, B. B., Trust, T., and Bond, M. A. 2020. The difference between emergency remote teaching and online learning. Available at: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-andonline-learning

Karakose, T. 2021. Emergency remote teaching due to COVID-19 pandemic and potential risks for socioeconomically disadvantaged students in higher education. *Educational Process: international journal 10*(3): 53-62. https://dx.doi.org/10.22521/edupij.2021.103.4

Karakose, T. 2021b. The impact of the COVID-19 epidemic on higher education: Opportunities and implications for policy and practice. *Educational Process: International Journal 10*(1): 7-12. https://doi.org/10.22521/edupij.2021.101.1

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … and 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

Landis, J.R., and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet Digital Health* 5(3): e105-e106.

Lund, B. D., and Wang, T. 2023. Chatting about ChatGPT: How may AI and GPT impact academia and libraries?. *Library Hi Tech News*, ahed-of-print. https://doi.org/10.1108/LHTN-01-2023-0009

OpenAI (2023). *GPT-4 technical report*. Available at: https://cdn.openai.com/papers/gpt-4.pdf

Sabzalieva, E., and Valentini, A. 2023. *Chat GPT and artificial intelligence in higher education: A quick start guide.* France: UNESCO

Sallam, M. 2023. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* (11)6. https://doi.org/10.3390/healthcare11060887

Smith, W. C. 2021. Consequences of school closure on access to education: Lessons from the 2013–2016 Ebola pandemic. *International Review of Education* 67(1-2): 53-78.

Stokel-Walker, C., and Van Noorden, R. 2023. What ChatGPT and generative AI mean for science. *Nature* 614: 214–216.

Strubell, E., Ganesh, A., and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 57: 3645-3650.

Sun, S. 2011. Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology* 11: 145-163. http://doi.org/10.1007/s10742-011-0077-3

Taecharungroj, V. 2023. "What Can ChatGPT Do?"; Analyzing early reactions to the innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing* 7: 35-49.

van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L. 2023. ChatGPT: Five priorities for research. *Nature* 614: 224–226.

Vieira, S. M., Kaymak, U., and Sousa, J. M. 2010, July. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems* (pp. 1-8), July, 18-23, 2010 - CCIB, Barcelona, Spain.

Wang, S., Scells, H., Koopman, B., and Zuccon, G. 2023. Can ChatGPT write a good Boolean query for systematic review literature search?. *arXiv* 2302.03495.

Warner, J., 2023. Biblioracle: Will artificial intelligence like ChatGPT bring the end for all writers? Chicago Tribune (Jan 07, 2023).

Warrens, M. J. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy* 5(4): 1. http://dx.doi.org/10.4172/2161-0487.1000197

Whittle, C., Tiwari, S., Yan, S., and Williams, J. 2020. Emergency remote teaching environment: A conceptual framework for responsive online teaching in crises. *Information and Learning Sciences* 121(5/6): 311-319.

ABOUT THE CONTRIBUTORS

Tijen Tülübaş is an Associate Professor of Educational Administration at Dumlupınar University, Kutahya, Türkiye. Her research interests include organizational behavior, higher education management, leadership, culture and identity. She has published numerous articles in leading international journals, authored a book and three book chapters on education/management.
Email: tijen.tulubas@dpu.edu.tr
ORCID ID: http://orcid.org/0000-0001-9406-8361

Murat Demirkol is Lecturer at the Vocational High School at Firat University (Turkey). His main fields of study are mentoring, educational leadership and administration, school administration and classroom management.
E-mail: mdkol@hotmail.com
ORCID ID: https://orcid.org/0000-0003-3108-3219

Tuncay Yavuz Ozdemir, PhD, is Associate Professor of Educational Administration in Firat University: Elazığ, TR. His main research interests include educational administration, leadership, educational leadership, and mentoring and research methods. He has published extensively in leading international journals and also authored books and chapters.
Email: tyozdemir@firat.edu.tr
ORCID ID: https://orcid.org/0000-0002-5361-7261

Hakan Polat is an Assistant Professor at the Department of Educational Sciences at Firat University. His main research interests are educational administration, educational leadership, school administration, classroom management, and teacher training. He has written articles on educational administration published in international journals.
Email: hakanpolat@firat.edu.tr
ORCID ID: http://orcid.org/0000-0002-0271-3747

Turgut Karakose is a Professor and Head of the Department of Educational Sciences at Dumlupinar University: Kutahya, Türkiye. His main research interests include educational leadership and management, higher education, psychology, and human behavior. He has published extensively in leading international journals and also authored books and chapters on education/management.
E-mail: turgut.karakose@dpu.edu.tr
ORCID ID: https://orcid.org/0000-0003-0346-8154

Ramazan Yirci, PhD, works an associate professor of Educational Administration at Kahramanmaras Sutcu Imam University, Türkiye. His main research interests include educational leadership and administration, mentoring and coaching, teacher induction, training school principals, educational psychology. His recent publications have focused on the effects of COVID-19 in education.
Email: yirci@ksu.edu.tr
ORCID ID: https://orcid.org/0000-0003-4696-7420