INTERNATIONAL JOURNAL OF MODERN EDUCATION STUDIES

| Volume 8 - No 1 - December 2024 - ISSN 2618-6209 |

Research

From “Can AI think?” to “Can AI help thinking deeper?”: Is use of Chat GPT in higher education a tool of transformation or fraud?

Yalçın Dilekli [1] **Serkan Boyraz** [2]

Article Type
Original Research

International Journal of Modern Education Studies
2024
Volume 8, No 1
Pages: 49-71
http://www.ijonmes.net

Article Info:

Received : 06.11.2023
Revision : 08.03.2024
Accepted : 20.03.2024

Abstract:

This research was conducted to see if using ChatGPT prompts students to think more deeply through reflection reports. The case study method and qualitative research methodology were used to carry out this study. Five graduate students in the Curriculum and Instruction department at Aksaray University's Social Sciences Institute who were teachers in various subjects and employed at various state school levels participated in the study. It was found that the majority of participants accepted all of the information presented by ChatGPT based on a citation as true, did not feel the need to control data reliability, and could be manipulated by ChatGPT while doing self-evaluation. Additionally, despite the fact that they prepared reflective reports in which they compared their essays with ChatGPT and included questions that prompted them to think critically and reflectively, as well as the fact that they had taken a graduate-level course on the teaching of higher order thinking skills, it was acknowledged that they could not demonstrate the expected performance in using higher order thinking skills other than to a limited extent. The onus should be on educators to pioneer positive examples of how to utilize ChatGPT and provide direction on how to harness its potential, supported by critical thinking, rather than to avoid using it and identify it as a tool to be avoided.

Keywords: Artificial intelligence, critical thinking, ethical considerations, reflective thinking, graduate students

Citation:
Dilekli, Y., Boyraz, S. (2024). From “Can AI think?” to “Can ai help thinking deeper?”: Is use of chat gpt in higher education a tool of transformation or fraud. *International Journal of Modern Education Studies, 8*(1), 49-71. https://doi.org/10.51383/ijonmes.2024.316

---

[1] Assoc. Prof. Dr., Aksaray University, Education Faculty, Aksaray, Türkiye. yalcindilekli@aksaray.edu.tr, http://orcid.org/0000-0003-0264-0231

[2] Assist. Prof. Dr., Aksaray University, Education Faculty, Aksaray, Türkiye. serkanboyraz@aksaray.edu.tr, http://orcid.org/0000-0001-6483-1397

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original authors and source are credited.

INTRODUCTION

Artificial intelligence (AI) is the replication of human intelligence in devices that have been designed to reason and acquire knowledge similarly to humans. AI aims to develop machines that are capable of carrying out operations that ordinarily require human intellect, such as comprehending natural language, spotting patterns, resolving issues, and making judgments (Lebovitz et. al., 2021; Turing, 2009). In 1950, Turing proposed the Turing Test, a standard for judging a machine's capacity for intelligent activity. AI systems centered on symbolic AI in the 1950s, using knowledge manipulation and symbolic representation. Expert systems were rule-based systems created by researchers in the 1960s that could solve particular issues by imitating the judgment of human experts (Chung & Silver, 1992). AI is now incorporated into many facets of modern life, from virtual assistants and recommendation tools to self-driving cars and cutting-edge medical software. The development of AI is still being driven by ongoing research and interdisciplinary cooperation (Bisconti et al., 2023), which shapes its potential applications and societal effects. One of the disciplines that seeks to use ChatGPT more effectively is education.

Higher education has recently received a lot of interest about the use of artificial intelligence (AI). AI-powered chatbots, like OpenAI's GPT-3, 3,5, and 4, have become an important resource in this situation. ChatGPT's capacity to increase student involvement is one of its main benefits in higher education. Students can ask questions and get prompt answers thanks to chatbots that enable real-time interaction. This quick feedback encourages active learning and may boost student involvement (Cotton et al., 2023; Rawas, 2023). Students can receive individualized learning support from ChatGPT. Chatbots can provide specialized recommendations and resources by examining students' questions and learning trends. This personalized method accommodates various learning preferences and fosters a more welcoming learning atmosphere (Adiguzel et. al., 2023; Fuchs, 2023).

ChatGPT, however, may unintentionally produce and disseminate false information. This could lead students’ astray in learning environments. If users are not attentive in how they interpret the material produced, it could result in incorrect conclusions in their research (De Angelis et al., 2023). What is more, students' ability to use thinking skills, especially critical thinking, may suffer from an over-reliance on ChatGPT. Genuine learning and intellectual growth can be hampered by relying on AI for solutions without comprehending the underlying concepts (Fuchs, 2023; Iskender, 2023).

At this point, another question arises: what do thinking skills refer? Although there are many definitions of the components of thinking skills, they refer to the synthesis, analysis, and evaluation levels of Bloom’s cognitive taxonomy (Swartz & McGuinness, 2004). These skills can be listed as searching for meaning (analytical thinking), creative thinking, critical thinking, decision-making, and problem solving (Swartz & Parks, 2004).

Swartz & McGuinness (2004) have added one more skill, meta-cognition. Reflective thinking is also accepted as a higher-order thinking skill (Schön, 1992). Meta-cognition is employed in each of the thinking skills. In fact, all these skills finally reach problem solving or decision-making (Costa, 1985).

In this context, this study aimed to determine whether the use of ChatGPT in higher education can mediate learners' use of higher-order thinking skills. For this purpose, the participants who took the Teaching Thinking Skills course at the graduate level and had a background in theoretical knowledge and practices in the use of higher-order thinking skills were asked to evaluate the texts created by ChatGPT from a reflective perspective. Details about the research procedure are given in the following part. Based on this reflective perspective, it was aimed at enabling students to think more deeply by using the skills of control data reliability, self-evaluation, and decision-making, another thinking skill within the scope of critical thinking, which are among the higher-order thinking skills, while evaluating the text created by ChatGPT.

METHOD

Model

This study, which aims to deeply examine the participants' use of higher-order thinking skills through the use of ChatGPT in higher education, was conducted through qualitative research methodology and the case study method. The case study approach is used to study the case of an individual, group, occasion, organization, or society and assists in giving a thorough understanding of the nature, procedure, or phenomenon of a particular example under examination by using a variety of data collection techniques, including interviews, observation, documents, and questionnaires (Kumar et al., 2022). Accordingly, detailed information about the participants of the study, the participant selection process, the creation of data collection tools, the research and data collection process, and the analysis of the data are presented in the following sections.

Participants

The participants of the research included five master’s students (see Table-1) studying at the Curriculum and Instruction program of the Social Sciences Institute at Aksaray University who were teachers from different fields and working at different levels of state schools. The selection of participants was based on two criteria. Firstly, they were required to take the Teaching Thinking Skills course offered in the given program in the 2022-2023 spring term. The course lasts fourteen weeks. The main topics covered are: What is thinking? Can thinking be taught? Why is teaching thinking necessary?; approaches to teaching thinking; higher-order thinking skills and different classifications; analytical thinking; critical thinking; creative thinking; reflective thinking; metacognitive thinking; decision-making; problem solving; relationships between types of thinking. This was the first criterion since the research was shaped upon whether ChatGPT could be used to provide activation of higher-order thinking skills such as reflective and metacognitive thinking, and these skills were the main topics of the given course. There were eight graduate students receiving this course. The second criterion was voluntary participation, and five out of eight students declared they would take part in the research.

Table-1

Information about Participants

	Age	Gender	Field	Works at
P1	40	M	Classroom	Primary school
P2	36	F	Computer	Secondary school
P3	32	F	Classroom	Primary school
P4	34	F	Philosophy	Secondary school
P5	41	F	Counseling and Psychological Guidance	High School

As can be seen in Table-1, of the five participants who voluntarily participated in the study, one was male and the other four were female. The average age of the participants was 36.6 years. Two of the participants work in primary school, two in middle school, and one in high school.

Data Collection Tools

There were three data collection tools used in the research. The main tool that shaped the research was participants' reflective papers, in which they compared their essays with ChatGPT's. The reflective papers included answers to four questions, two of which were directly about ChatGPT's performance in writing essays and were intended to require students to compare their essays with ChatGPT's. These questions were:

1. What do you think about ChatGPT's essays?
    a) Do you agree with the ideas put forward in the essay you reviewed? Why?
    b) Are there any ideas you disagree with in the essay you reviewed? Why?
    c) When you compared this essay with your own, did you find it scientifically better or worse than yours? Why?
    d) Are there any shortcomings in the essay you reviewed?
2. Is there any information and/or opinion in the essay you reviewed that you would like to add to your own essay?

The remaining two questions were intended to reveal participants' personal attitudes towards the use of ChatGPT for academic purposes, together with any ethical considerations they may have. These questions were:

3. The essay you reviewed had been prepared by the ChatGPT application, which used an artificial intelligence algorithm. Do you think it is ethically appropriate to use such practices in scientific studies? Why?
4. Would you use ChatGPT in your future scientific studies (preparing assignments and essays, writing a thesis, etc.)? Why?

The two other data sources were the two essays written by the participants and ChatGPT. The first topic was given as “Can you explain similarities and differences between reflective and metacognitive thinking?” which would be placed in the analysis level of Bloom’s cognitive taxonomy. At this level, students are expected to detect and compare relationships, and it is advised that the content be familiar to them (Bloom, 1956). As participants had already studied these two terms in the Teaching Thinking Skills course, familiarity with the content was provided. The second topic was assigned as “What kind of similarities and differences are there between the sub-skills put forward by Facione (2011) and Swarts and Parks (1994) regarding critical thinking? Which one do you think is more logical? Why?”. Different from the first essay topic, this topic had two parts, first of which (What kind of similarities and differences are there between the sub-skills put forward by Facione (2011) and Swarts and Parks (1994) regarding critical thinking?) would be placed at analysis level again, but second (Which one do you think is more logical? Why?) would go to the evaluation level of Bloom’s cognitive taxonomy. The evaluation level includes the act of passing judgment on an object's worth in relation to ideas, works, solutions, techniques, materials, etc. and entails applying standards and criteria to determine how precise, efficient, cost-effective, or satisfying certain details are (Bloom, 1956).

Data Collection Procedure

Data were collected through three data collection tools. First, the essays prepared by ChatGPT were analyzed from the perspective of students and research. Then, the essays prepared by ChatGPT, and students were analyzed and compared with each other by the researchers. Finally, the reflective reports of the participants were analyzed by the researchers. The data was collected between 15/07/2023 and 15/08/2023 through e-mails. The participants were sent the essay topics and given three weeks to prepare them. When they sent their essays to researchers through e-mails, the essays prepared on the same topics by ChatGPT were sent to participants and they were asked to examine them and return their reflective papers in a week. All five participants complied with the time limits and delivered the requested data to the researchers.

Data Analysis

The first part of the data analysis included assessing essays prepared by both students and ChatGPT through a rubric that included four main criteria and was prepared by the researchers. The first criterion was the accuracy of information, and essays were examined in terms of providing valid information. The second criterion was the coherence of the text, which pointed out unity in terms of meaning and appropriate ties among sentences and was evaluated on four issues (Johns, 1986), namely relations between paragraphs, use of linking words, focusing on one idea in each paragraph, and indicating a main sentence. The third criterion was the level of correct answers, an indication to what extent participants provided the required information. Finally, the last criterion was the use of correct references, which was examined through validating if the cited resources included the given information. In other words, all citations were carefully examined in the original resources by researchers. The rubric was subjected to three expert opinions, two of whom holding a PhD in Curriculum and Instruction and one in Counseling and Psychological Guidance.

When the final form of the essay was formed after experts' suggestions, all essays were rated as good, fair, or unsatisfactory in terms of these criteria independently by researchers, and the consistency between ratings was examined through Cohen's kappa, which is commonly used for quantifying inter-rater agreement on a nominal scale (Warrens, 2015), like the one used in this research. So, a total of 12 essays -ten coming from 5 participants and 2 coming from ChatGPT- were rated independently, and Cohen's kappa was calculated for all of them. Cohen's kappa values ranged between .80 and .91 which could be interpreted as substantial and almost perfect agreement (McHugh, 2012).

The second part of the data analysis included a content analysis of the reflection papers prepared by participants. The content analysis was carried out as a collaborative data analysis procedure in which two or more researchers work together to focus on and communicate about a common set of data in order to arrive at a shared interpretation, as collaboration promotes systematicity, clarity, and transparency (Cornish et al., 2014) that are crucial in terms of inter-coder reliability in qualitative research like this. As a result, the content analysis went on until an agreement was reached by the researchers.

Research Procedure

In the first place, researchers sent essay topics to participants, and then the participants prepared two essays with the given topics and sent them to the researchers. In the third step, researchers asked ChatGPT to prepare two essays with the same topics. It is to be noted that researchers first used ChatGPT's unlicensed version (3.5); however, its performance was low in terms of providing information and using in-text references. Then, a ChatGPT (version 4.0) license was bought, and the essays provided by this version were used in the research. The essay topics were prompted in ChatGPT as they were given to participants only with "adding required citations and references" at the end. The essays prepared in this way were sent to participants.

Then came the fourth and fifth steps of the research, in which participants examined essays prepared by ChatGPT through reflection questions that were provided by researchers and explained in the Data Collection Tools part. Participants were directed to compare their essays and performances with ChatGPT's with these reflection questions. The primary aim of generating such a reflection process was to activate higher-order thinking skills of participants. As a result, participants sent two reflection papers, one for each essay, to the researchers.

In the sixth and seventh steps, essays prepared by participants and ChatGPT were evaluated by researchers using the rubric that was explained in the Data Analysis part. Each essay was separately examined in terms of the criteria in the rubric. In the eighth step, participants' and ChatGPT's rubric evaluations were compared, which provided an overview of and opportunity to compare human and artificial intelligence performances on given tasks. This step led researchers to be able to comment on the reflections of participants.

Finally, in the ninth step, participants’ reflection papers were subjected to content analysis that aimed to investigate if students followed critical and metacognitive thinking skills they studied in the Teaching Thinking Skills course, and whether they were reflective.

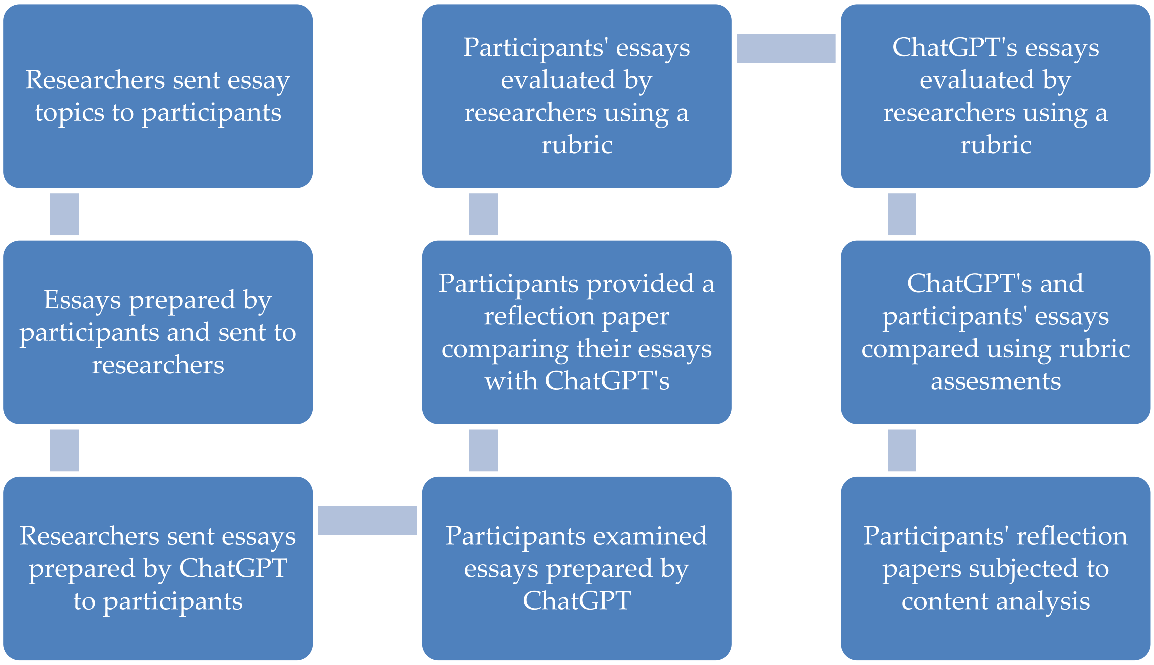
Researchers sent essay topics to participants
Essays prepared by participants and sent to researchers
Researchers sent essays prepared by ChatGPT to participants
Participants examined essays prepared by ChatGPT
Participants provided a reflection paper comparing their essays with ChatGPT's
Participants' essays evaluated by researchers using a rubric
ChatGPT's essays evaluated by researchers using a rubric
ChatGPT's and participants' essays compared using rubric assesments
Participants' reflection papers subjected to content analysis

Figure-1 Research Procedure

In alignment with the overarching commitment to ethics, this study stringently adhered to all provisions delineated in the "Higher Education Institutions Scientific Research and Publication Ethics Directive." It is imperative to note that there were zero instances of activities which might infringe upon the clauses stated under the "Actions Against Scientific Research and Publication Ethics."

Ethical Review Board: Aksaray University Ethical Review Board

Date of Ethics Review Decision: 20.06.2023

Ethics Assessment Document Issue Number: 2023/04-12

RESULTS

Research Question-1: Is ChatGPT an effective source for academic research/studies? Is it a reliable information source? Does it represent the information coherently? Is the information provided by ChatGPT sufficient? Does ChatGPT use the correct references?

With the aim of replying to this question, essays prepared by ChatGPT (4.0) that included answers to the two questions given below were evaluated according to the rubric. The first essay provided by Chat GPT is given below:

Essay Question-1: Can you explain similarities and differences between reflective and metacognitive thinking adding required citations and references?

> *Reflective thinking and metacognition both focus on the abilities to monitor and evaluate one's own thought processes. These two concepts have overlapped aspects and some distinctive features.*
>
> *Similarities:*
>
> *Both involve the individual being aware of his or her own thought processes and managing these processes (Flavell, 1979).*
>
> *Reflective thinking and metacognition enable individuals to be more effective in problem solving, learning and decision-making (Zohar & Barzilai, 2013).*
>
> *Differences:*
>
> *Metacognition is generally a broader concept and encompasses an individual's ability to understand and control their own mental processes and strategies. On the other hand, reflective thinking generally refers to the individual's process of evaluating a particular situation or experience (Schön, 1983).*
>
> *While metacognition focuses more on learning and problem-solving strategies, reflective thinking focuses more on evaluating and understanding learning experiences and results (Kitchener, 1983).*
>
> *While metacognition is often defined as "thinking about thinking," reflective thinking generally refers to reflecting on a situation or experience and is often performed to improve a particular action or situation (Schön, 1983).*
>
> *References:*
>
> *Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American psychologist, 34(10), 906.*
>
> *Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: Current and future directions. Studies in Science Education, 49(2), 121-169.*
>
> *Schön, D. A. (1983). The reflective practitioner: How professionals think in action (Vol. 5126). Basic books.*
>
> *Kitchener, K. S. (1983). Cognition, metacognition, and epistemic cognition. Human Development, 26(4), 222-232.*

The first essay topic was placed at the analysis level of Bloom's taxonomy. When the ChatGPT (CG) essay for this question was examined in terms of the accuracy of the given information, it was found that its performance was good as two main similarities and three differences explained in the literature were given. The second criterion in assessment was defined as the coherence of the text. The CG essay lacked relation between paragraphs as it provided most of the content in the form of a so-called "bullet point". The essay, as a result, was evaluated as unsatisfactory in this aspect. On the other hand, the essay was categorized as good in terms of use of linking words within paragraphs such as "both, on the other hand, etc.". When it comes to focusing on one idea in each paragraph, the essay was again assessed as good. The essay started with an introductory explanation, including the main idea, so it was scored as good regarding this aspect. The level of giving correct answers was found to be good because the CG provided all required answers. The last criterion in the rubric was use of thecorrect references. The CG provided four references in the essay, as can be seen above. When the two references, Flavell (1979) and Zohar and Barzilai (2013), were analyzed, the given information was verified. Yet, the information cited from Schön (1983) by ChatGPT could be partly authenticated. When it comes to the fourth reference, Kitchener (1983), the cited information could be found partly again since the cited work did not include the topic "reflective thinking." So, the overall performance in terms of using correct references was unsatisfactory. The second essay provided by Chat GPT is given below:

Essay Question-2: What are the similarities and differences between the sub-skills that Facione (2011) and Swartz and Parks (1994) put forward regarding critical thinking? Which one is more logical? Why? Can you explain these by adding required citations and references?

> *Facione (2011) evaluates critical thinking through seven key skills or sub-skills:*
>
> *Interpreting: Understanding and explaining communication.*
> *Analysis: Determining the relationships of claims and arguments.*
> *Evaluation: Determining the strength of an argument.*
> *Inference: Identifying conclusions and constructing arguments.*
> *Explanation: Expressing results and experiences clearly.*
> *Self-correction: Identifying and correcting mistakes and biases.*
> *Cognitive integrity / Coginitive ethics: Evaluating and appreciating your own thoughts and those of others.*
>
> *On the other hand, Swartz and Parks (1994) conceptualize critical thinking as a set of different "thinking tools" and focus on teaching these tools across a variety of disciplines and contents. These tools usually include:*
>
> *Comparison and classification.*
> *Exemplification.*
> *Analysis of contrasts and anomalies.*
> *Analysis of sensitivity and assumptions.*
> *Cause and effect relationships.*
> *Evaluation and judgment.*
> *Creating hypotheses.*
>
> *Critical thinking, in both approaches, includes sub-skills such as analysis, evaluation, inference and explanation. However, while Swartz and Parks' approach focuses more on thinking processes, Facione's approach focuses more on thinking abilities.*
>
> *Determining which approach "makes more sense" often depends on the instructional purpose, the learning context, and the specific needs of the student. Facione's approach is perhaps more general and comprehensive, but*

> *Swartz and Parks' approach is perhaps more concrete and useful for teaching and practice.*
> *References:*
> *Facione, P. A. (2011). Critical Thinking: What It Is and Why It Counts. Insight Assessment.*
> *Swartz, R. J., & Parks, S. (1994). Infusing the Teaching of Critical and Creative Thinking into Content Instruction. Pacific Grove, CA: Critical Thinking & Software.*

The second essay topic was placed at the analysis and evaluation levels of Bloom's taxonomy. The text provided by the CG for this essay was evaluated as unsatisfactory in terms of the accuracy of the given information. That is because the CG added one more dimension for critical thinking sub-skills, which was cognitive integrity/cognitive ethic, which was not verified from the cited reference (Facione, 2011) by the researchers. Similarly, the CG presented seven critical thinking sub-skills citing Swartz and Parks (1994), but five of them (Comparison and classification, Exemplification, Analysis of contrasts and anomalies, Analysis of sensitivity and assumptions, Evaluation and judgment) were given in the other types of thinking skills, such as analytical thinking, problem solving, and so on rather than critical thinking in the given reference. When it comes to the coherence of the text, it was seen that the text was mostly in the form of so-called bullet-point form, together with three short paragraphs. Besides, these paragraphs were not connected to each other. In this respect, the text was found to be unsatisfactory. On the other hand, sentences within the same paragraph were linked to each other through linking words that showed contrasts and cause-and-results relationships. Therefore, the text was rated as good in this respect. The text was rated good again in terms of focusing on one idea in each paragraph. For example, the following paragraph directly focused on comparing two approaches to critical thinking as required by the essay topic:

> *Critical thinking, in both approaches, includes sub-skills such as analysis, evaluation, inference and explanation. However, while Swartz and Parks' approach focuses more on thinking processes, Facione's approach focuses more on thinking abilities.*

The text was evaluated to be good again in terms of indicating the main sentence as CG provided the main idea in the last paragraph of the text:

> *Determining which approach "makes more sense" often depends on the instructional purpose, the learning context, and the specific needs of the student. Facione's approach is perhaps more general and comprehensive, but Swartz and Parks' approach is perhaps more concrete and useful for teaching and practice.*

The level of giving correct answers was unsatisfactory because the expected number of items concerning sub-skills of critical thinking proposed by the two approaches by Facione (2011) and Swartz and Parks (1994) was 12, while the CG provided 14 items, only seven of which were correct. This second essay was not evaluated in terms of the last criterion in the rubric, which was the use of correct references, as the essay topic already limited the content to two references.

Research Question-2: How was the performance of participants compared to ChatGPT's?

Participant-1's (P1) essay was evaluated using the same rubric, and it was seen that the participant explained two main similarities and three main differences, so the performance was rated as good in terms of the accuracy of the given information. When it comes to the coherence of the text criterion, P1 essay was evaluated as good in terms of relation between paragraphs, use of linking words within paragraphs, and indicating a main sentence. However, it was rated as fair in terms of focusing on one idea in each paragraph since there was too much repetition in different paragraphs. For example, P1 started three paragraphs with these sentences below, all of which indicating a very similar idea:

Paragraph-1: "*Reflective and metacognitive thinking are interrelated thinking skills…*"

Paragraph-5: "*Reflective and metacognitive thinking are complementary skills…*"

Paragraph-6: "*Reflective and metacognitive thinking are not apart, they are together…*"

Since the essay started with an introductory explanation including the main idea, it was scored as good regarding this aspect. The level of giving correct answers was found to be good as P1 provided all required answers. Since there were not any citations within the text or references in the end, P1's performance in terms of using correct references was found to be unsatisfactory. As a result, P1 had a similar performance with CG in terms of the accuracy of the information and level of giving correct answers, while P1 had a better performance in terms of the coherence of the text and CG had a better performance in terms of using correct references.

Participant-2's (P2) essay was evaluated as fair in terms of the accuracy of the information since there was a misconception in terms of the two skills compared: "*The fine detail that separates reflective thinking and metacognition skills is; Metacognition is the cognitive process of the individual based on the pros and cons of the individual's learning; Reflective thinking is; We can say that it is the individual's ability to evaluate himself in general and realize what is suitable for him based on his pros and cons.*" P2's performance in terms of the coherence of the text was good in all sub-dimensions because relations were provided between paragraphs, linking words were used within paragraphs, each paragraph focused on one main idea, and the essay started with a main sentence. The essay was evaluated as unsatisfactory in terms of the level of giving correct answers since there was a misconception of terms and inappropriate examples were given: "*…As a result of a student realizing that his desire to study decreases when he spends too much time on the internet; He makes study plans by limiting the time he spends on the internet (Reflective thinking)…*" This example is more suitable for self-evaluation sub-dimension of critical thinking. The use of correct references was unsatisfactory as there was only one in-text and there was not a references part at the end of the text. As a result, P2 had a better performance only in the coherence of the text criterion, while in all other three criteria CG showed a better performance.

Participant-3'S (P3) essay was assessed as good in terms of the accuracy of the information. When it comes to the coherence of the text, the essay was found fair in terms of focusing on one idea in each paragraph and relations between paragraphs because some paragraphs indicated both similarities and differences, and none of the paragraphs were linked to previous or coming paragraphs. For example, while second paragraph explained similarities between the two terms, third paragraph explained differences between them, and the fourth paragraph contained both similarities and differences. P2's performance was good in terms of using linking words within paragraphs and indicating a main sentence. The level of giving correct answers was good. On the other hand, there weren't any in-text citations or references at the end, so the performance in terms of using correct references was unsatisfactory. As a result, P3 performed better in terms of the coherence of the text while the performances of P3 and CG were the same in terms of the accuracy of the information and the level of giving correct answers. CG performed better than P3 in terms of using correct references.

Participant-4's (P4) essay was found to be good in terms of the accuracy of the information, as all the information provided by the participant was correct. The essay was evaluated as good in terms of the coherence of the text in all its sub-dimensions. The level of giving correct answers was evaluated as fair since one item was missing in both similarities and differences dimensions. When it comes to the last criterion, which is the use of correct references, it was assessed as unsatisfactory as there was only one in-text citation, which was not provided as a reference part at the end of the text. As a result, P4 performed the same as CG in terms of the accuracy of the information and was better than it when it comes to the coherence of the text. However, ChatGPT performed better than P4 in terms of the level of giving correct answers and using correct references.

Participant-5's (P5) essay was evaluated as good in terms of the accuracy of the information. As P5 provided similarities and differences in the form of bullet-points separately at first and then explained them in detail paragraph by paragraph, the assessment for the coherence of the text was good in all sub-dimensions. P5's performance in terms of the level of giving correct answers was good again. The use of correct references, on the other hand, was unsatisfactory since there weren't any in-text citations or a references part at the end of the text. As a result, P5 performed the same as ChatGPT in terms of the accuracy of the information and the level of giving correct answers. On the other hand, P5 performed better than ChatGPT in terms of the coherence of the text and worse than it in terms of using correct references.

P1's essay-two performance was evaluated, and it was found that P1's performance in terms of the accuracy of the given information was unsatisfactory because P1 explained Facione's classification of critical thinking sub-skills while Swartz and Parks' classification was not provided correctly. There was a limited relationship between paragraphs; various ideas were placed within one paragraph; there was a limited use of linking words within paragraphs, and there was not a clearly stated main idea; so, the coherence of the text was evaluated as unsatisfactory. Since half of the information provided by P1 was not accurate, P1's performance in terms of the level of giving correct answers was unsatisfactory again. It is to be noted here that P1 had a personal judgment as required by the second part of the essay topic. As a result, ChatGPT outperformed P1 in terms of the coherence of the text, but they performed equally in terms of the accuracy of the given information and the level of giving correct answers.

P2's performance in terms of the accuracy of the given information was good as P2 provided and matched all sub-dimensions explained by two resources. The overall rating in terms of the coherence of the text was good since P2 set relations between paragraphs, used conjunctions within paragraphs, focused on one idea in each paragraph, and indicated a main sentence. Similarly, P2 was rated as good in terms of the level of giving correct answers. As a result, P2 performed better than ChatGPT in all aspects of the evaluation.

When P3's essay two was examined, it was seen that P3 explained nine sub-dimensions, while 12 was expected. So P3's performance in terms of the accuracy of the given information was good. The coherence of the text, similarly, was good since relations between paragraphs were provided, in-paragraph links were set, paragraphs included only one idea, and a main idea was indicated. As three sub-dimensions were missing in the essay, the rating of the level of giving correct answers was fair. To conclude, P3 performed better than ChatGPT in all three aspects.

P4's performance in terms of the accuracy of the given information was good because P4 provided all twelve sub-dimensions as stated in the two resources. The text's coherence was also strong since there were connections between paragraphs, interconnections within paragraphs, only one thought per paragraph, and indications of the core topic. The level of giving correct answers was also good, as all sub-dimensions were explained correctly. As a result, P4 had a better performance than CG for all criteria.

P5's performance in terms of the accuracy of the given information was fair since Facione (2011)'s three and Swartz and Parks (1994)'s two sub-dimensions were missing. On the other hand, the given sub-dimensions were matched. The assessment of the text's coherence was good in all sub-dimensions since P5 initially listed the similarities and contrasts in distinct bullet points before going into more depth about each one in a paragraph. Also, there were many linking words that increased the readability. When it comes to the level of giving correct answers, it was rated as fair since P5 added two sub-dimensions, namely curiosity and being systematical, which couldn't be verified in the given resources. To conclude, P5 performed better than CG in all aspects.

Research Question-3: Did the participants use critical thinking skills while writing reflection papers?

Self-reflection Reports for Essay-1

P1 indicated in the reflection report that he agreed with the content provided by CG, saying that "*… The explanations stated in the essay helped us to understand better the similarities and differences between reflective and metacognitive thinking skills…*" Besides, he took the information and citations provided by CG for granted as he clearly expressed: "*…I agree with the whole content presented in this essay.*" On the other hand, P1 pointed out that his essay was better than CG’s because P1 did not only explained similarities and differences but also examined relations between the two types of thinking skills: “*These two skills are intertwined so it is impossible to explain these terms without mentioning their relationship. In my essay, I mentioned about this relationship, but ChatGPT did not…*” P1, according to the reflection paper, recognized his own scientific deficiencies while examining CG’s essay by expressing: “*…Metacognition is a wider term than reflection and I would set my essay over this…*”

According to her reflection paper, P2 conceded most of the content provided by CG. The point that she disagreed was expressed as: “*I disagree the idea that metacognitive thinking is a wider term than reflective thinking which was claimed by ChatGPT.*” P2 found her essay scientifically better than CG’s and indicated it as follows: “*Mine is better because I made inferences by searching for information in a long while. I presented it in my essay by synthesizing them.*” Similarly, P2 found CG’s deficient as it lacked coherence in presenting the content and providing inferences: “*The content in the essay composed of different parts which are not related to each other, and it does not reach any results.*” P2 clearly indicated that she did not want to take anything from CG’s essay as she found many deficiencies in that work.

P3 agreed with the ideas provided in CG’s essay indicating that “*I especially agree with the ideas claimed related to metacognition as one of the mostly discussed topics in philosophy is thinking about thinking… Since reflective thinking helps students deciding positive and negative sides of learning experiences, I agree with the ideas proposed in this essay.*” P3 does not find CG’s essay better than hers and explained it as: “*I prepared my essay by analyzing many resource books; however, there are only 3 references in ChatGPT’s essay.*” According P3 the main problem in CG’s essay is the lack of explanation and presenting the content in the form of bullet points: “*The essay consisted of bullet points, and it lacks coherence.*” P3 stated that she would add the following idea stated by CG to her essay: “*Reflective thinking and metacognition provides individuals with being more active in problem solving, learning and decision making.*”

P4 admitted most of the ideas indicated by CG by indicating that she wrote similar things as provided in CG’s essay. There was a point that she disagreed, and she explained it as follows: “*I don’t agree that metacognitive thinking covers reflective thinking…*” P4 found CG’s essay better than hers due to citations given by CG. When it comes to the deficiencies of the essay, P4 stated that there weren’t any concrete examples while explaining the issue: “*There is no interpretation and examples in ChatGPT’s work…*” Finally, P4 explained that there wasn’t anything she wanted to transfer to her essay, and she explained it as follows: “I read a lot of information and made interpretations while preparing my essay; however, there is not interpretation in CG’s essay.

P5 stated that she acknowledged all of the ideas indicated in CG’s essay. On the other hand, she found her essay scientifically better than CG’s claiming that: “*I research the topic in many various sources and gave examples and details in order to increase understandability; however, in ChatGPT’s essay there were many short definitions only without any examples…*”

Self-reflection Reports for Essay-2

According to the reflection paper of P1 for essay two, P1 admitted all the claims by CG except that “*ChatGPT indicated that Facione’s approach was overwhelming than Swartz and*

Park's, but I think it depends on the teaching approach…". P1 stated that his and CG's essays were scientifically equal. On the other hand, P1 explained that CG didn't explain mutual and different points of the two approaches. P1 declared that he would like to add his essay two points from CG's: "*…I would like to add cognitive integrity and ethic sub-dimensions to my essay…*"

P2 indicated that she agreed most of the content proposed by CG; however, she stated that she couldn't find cognitive integrity and ethic sub-dimensions in the proposed resources. She found her own performance in essay two better than CG because of the lack of coherence and understandability in its essay. Furthermore, CG's essay was criticized as being in the form of copy-paste taken from search engines. P2 surprisingly wanted to add her essay cognitive integrity and ethic sub-dimensions as they were indicated in CG.

P3 stated that she agreed with the ideas claimed in CG's essay as there were many mutual points between her and its essay. Yet, there were also ideas that P3 didn't agree on: "*…I am on behalf of Swartz and Park's classification but ChatGPT does not support one of them clearly…*" P3 indicated that CG's essay would be better than hers in terms of scientific aspect as CG would use documents in different languages other than Turkish: "*…ChatGPT's essay may be more scientific than mine because I have no opportunity to check academic papers in other languages than Turkish…*" Yet, P3 evaluated CG's essay as insufficient in terms of detailed explanation and exemplification. She added that she would like to transfer following sentence into her essay: "*…Since I couldn't recognize that detail, I would like to add the following comment into my essay:* "Both approaches include sub-skills such as critical thinking, analysis, evaluation, implication and explanation. *However, Swartz and Park's approach mostly focuses on thinking processes while Facione's focuses on thinking skills.*"

P4 explained that she agreed on the ideas stated by CG except those about Swartz and Park's as CG proposed more sub-dimensions than hers. According to P4, her essay was better than ChatGPT's and the main deficiencies of CG's essay were lack of coherence and inferences: "*I don't want to transfer anything to my essay from ChatGPT's as it all included copy and paste information…In this essay we were supposed to give our interpretations and inferences…*"

P5 explained that she agreed on all claims by CG: "*I agree everything as similarities and differences have been clearly stated…*" She found her essay better than CG's since she provided more examples. She indicated that she would add a part of the similarities and differences proposed by CG as she hadn't included them in hers and found her essay more limited in terms of this aspect: "*…I wrote limited number of similarities and differences compared to ChatGPT…*"

Research Question-4: Do the participants find using CG for academic purposes ethically appropriate? Are they willing to use it in future research?

According P1 using CG in academic purposes may be ethical; however, data reliability and privacy, accuracy of the algorithm, authenticity and humanitarian responsibility may be problems. P1 explains it as follows: "*…As a result, by taking into ethical concerns ChatGPT can be used in the research…*" P1 expressed that he would be willing to use ChatGPT only for literature review to save time in the future research.

According to P2, using CG for academic purposes is not ethical because it is composed of a "copy and paste" algorithm. She indicated "*It is an algorithm based on copy and paste, so it is not ethical to use it…*" So, she rejects to use CG in the future research by saying that "*Learning is a cognitive development and reaching the information from pre-ready resources does not provide any benefits in terms of cognitive development. You learn only when you read and research yourself.*"

P3 finds using CG in academic research ethical by indicating: "*I googled these two terms and found many articles. ChatGPT does the same and presented us by refining them…*" Besides, P3 explained open willingness to use CG in future research as she explains: "*Using ChatGPT saves time, especially when you review the literature…*"

According to P4, there are not any ethical considerations in using CG in academic works; however, it may produce stereotype works unless the content provided by CG reproduced by human beings: "*Stereotype works are obstacles in front of the development of science and using ChatGPT without any interference may not help scientific development…*" She stated that she would use CG in future research for literature review but also filtering it through her mind.

P5 does not find using CG for academic purposes ethical: "*The information that is not learned does not belong to the individual… I don't think that an essay prepared by AI would belong to learners… AI uses limited resources…*"

DISCUSSION

The first research question aimed to investigate ChatGPT's effectiveness in academic research. When ChatGPT's first essay on comparing reflective and metacognitive thinking was analyzed under the four criteria, it was seen that ChatGPT explained similarities and differences between the two terms correctly. As a result, ChatGPT's performance in providing information for a question at the analysis level of Bloom's cognitive taxonomy can be evaluated as successful. Besides, ChatGPT was good at presenting the content in the form of an essay in terms of using linking words within paragraphs, placing only one idea in each paragraph, and indicating a main sentence, while there was a lack of providing relations between the paragraphs. Although it was rated as fair in terms of coherence, it might be developed in this area. When it comes to the level of giving correct information, it was seen that ChatGPT performed mostly well. Yet, when the cited works were analyzed, some problems were found. Firstly, ChatGPT cited Flawell (1979) for the information it provided on reflective thinking; however, when this work was analyzed by the researchers, the term "reflective" was found only in one sentence, and that sentence does not include the information provided by ChatGPT. Secondly, the statement cited from Zohar and Barzalai (2013) could not be found in the given source, either. Thirdly, when the sources of Schön (1983) and Kitchener (1983) were examined, some of the statements presented by ChatGPT by referring to these sources could not be found. Through the reference to Schön (1983), definitions of the concepts of reflective and metacognition were made, but the terms metacognition and metacognitive were never used directly in the related book. Yet, the definitions were accurate. In this case, it can be concluded that ChatGPT may have obtained the relevant definitions not directly from this source but from secondary sources referring to it. According to Dwivedi et al. (2023), even though ChatGPT is very sophisticated, it is still an AI model that relies on correlations and patterns it has learned from training data to function, which means that the text that is created could include errors, prejudices, and other types of false information that could undermine the validity of academic study. What is more, papers with minor originality and faults that go unnoticed are becoming more and more common, and ChatGPT will exacerbate this issue and provide articles that are frequently right but often compelling. This might, after all, skew scientific evidence, encourage plagiarism, and disseminate false information. Then, it would be advised to warn learners about these issues and develop a critical perspective with them in terms of using ChatGPT for academic purposes.

The second essay topic is a question at both the analysis and synthesis levels. The CG was found unsatisfactory in terms of the accuracy of information and the correct answer rate criteria in this question because five of the fourteen skills defined as sub-dimensions of critical thinking by referring to Facione (2011) and Swartz and Parks (1994) could not be found in the relevant sources. The CG states that he found four of these skills in Swartz and Parks (1994) and one of them in the other source. Swartz and Parks (1994) is the source where the CG makes the most mistakes in the essay. The relevant work is a book and is not open access on the internet, so it is possible that information might be taken from open access published secondary sources that cite this source, and this may have caused the problem. There are a number of research in the literature pointing out the incomplete or outdated knowledge as a limitation of the CG (Hariri, 2023; Ray, 2023). In terms of coherence, the CG's performance was found to be fair because, except for the inter-paragraph relationship criterion, CG was able to express the information appropriately and, more importantly, finally reach a synthesis-level conclusion by making a judgment. Mitrović et al. (2023) also indicate that CG has the capacity to produce grammatically perfect and convincingly human responses to numerous inquiry types from various fields. Thereupon, learners might be directed to examine CG’s language use to improve and encouraged to ask CG’s decisions in complicated situations (academic work and research) to have a starting point for themselves rather than copying it.

When the participants' reflective reports on CG were analyzed, it was seen that all participants except one accepted the information provided through in-text citations as correct. In this respect, it can be said that these participants did not use the sub-skill of controlling data reliability, which is a part of critical thinking skills, despite having taken a graduate course on this subject. The only participant who was skeptical about the accuracy of the information said that she could not check the accuracy of the information by expressing the possibility that CG could search in different languages. The literature holds a debate about whether using CG has the potential to decrease critical thinking and creativity (Plebani, 2023; Sallam et al., 2023). If this is the case, then the onus should be on educators not to avoid the use of CG and label it as a tool to be avoided, but to pioneer good examples of how to use it and provide guidance on how to harness its potential, accompanied by critical thinking. In addition, it is seen that all participants stated that their essays were better in terms of creating the text from a holistic point of view, supporting the explanations by giving examples, and making connections within and between paragraphs. The reason why four of the five participants criticized CG for not giving concrete examples and not elaborating the narrative in terms of the coherence and comprehensibility of the narrative may be the habits they have acquired in their education so far. Because in the Turkish education system, the tradition of discussing the topic through examples is dominant, and participants wrote their essays by providing examples. However, CG did not place any examples in its two essays, which is not surprising as the prompts used did not include any demand to provide examples. In fact, the participants also expressed opinions about CG's essays such as whether they were clear, core, short or concise. Again, three out of five participants stated that they wanted to include the main idea sentences given by CG in their essays. In this respect, CG can be seen as adequate in making short and clear evaluations by presenting the main idea, which makes it an important source when learners lose their way.

A general evaluation of the participants' and CG's first essays in terms of the four criteria in the rubric reveals that CG was behind the participants only in terms of coherence, but in all other criteria, CG performed either equal to or better than the participants. In terms of the accuracy of the information, no participant outperformed CG. When the performances of the second essay were compared, it was observed that all participants, except one, outperformed CG in all criteria. The main reason for this difference between the performances of the participants and CG in the two essays may be that there were no resource limitations in the first essay and open access resources could be used. On the other hand, the sources that should be used for the second essay were specified in the question, and one of these sources (Swartz & Parks, 1994) is a book that is not open access, so it is likely that CG could not access it directly. As a result, CG created this essay based on limited and incomplete information and did not perform better than the participants. This indicates that using CG when resources that it could/should use are not open-access, its potential is limited and expectations should be kept low in terms of the performance. This can be an advantage when the aim is to limit learners' use of CG while it is a big disadvantage if learners need AI support due to a lack of time.

The participants performed similarly to CG in their first essay, but according to the self-assessment results in their reflective reports, they think that they performed better than CG because all participants except one stated that they did not have anything they wanted to take from CG's essay into their own essay. On the other hand, in the second essay, although all but one of the participants performed better than CG in all aspects, four of the participants stated that they wanted to take information, ideas, or insights from CG's essay. This may be because CG provided more information than the participants by giving citations, even if incorrectly. According to the comparisons made by the students, they accepted the extra items presented by CG, which were incorrect, as correct. From this point of view, the participants did not feel the need to check the reliability of the data, even though they had previously learned it in the course and prepared an essay about it. Accordingly, CG could not mediate the use of higher-order thinking skills, even though participants were directed through questions in reflection papers. It seems that these four participants performed self-evaluation, a sub-dimension of critical thinking, incorrectly. Only one of the participants stated that she did not have the opportunity to check the data reliability by expressing the possibility that the cited information provided by CG might be taken from the sources written in different languages, and therefore she did not want to transfer any information to her own study.

Three out of five participants did not see any ethical problem in the use of CG in academic studies, and these participants stated in their reflective reports that there were sections that they could transfer from CG's essay to their own essays. From this point of view, it can be said that there is consistency in the opinions of the participants in their reflective reports. These participants also stated that they would filter the information they would receive from CG through their own filters, which they did not do according to reflective reports. Accordingly, it can be concluded that these participants think that CG can be an important source of information, but they do not trust it completely. On the other hand, the other two participants do not consider the use of CG in academic research ethically appropriate and do not think that there can be a transfer of knowledge from CG's essay to their own essays. It is to be noted here that the literature is full of ethical considerations in using CG in the academic context in many fields (Liebrenz et al., 2023; Stahl & Eke, 2024). Since it would be impossible and, more importantly, unnecessary when the advantages it provides are considered to prohibit using CG in carrying out academic tasks, learners can be supported to use it in a more lecturer-controlled way in which ethical issues are eliminated. The lecturer-controlled way expressed in the previous sentence does not intend to highlight censoring but providing limitations in which learners will not get lost, save time and energy, and actively use higher order thinking skills, as the researchers of this article tried to reach. This might lead to CG or AI being a tool for transformation of education rather than fraud.

LIMITATIONS AND RECOMONDATIONS

The research is limited to the data obtained from the study group. In addition, since this is a qualitative study, it is difficult to reach generalizations. Another limitation is that critical, analytical, and reflective thinking, metacognition, and decision-making skills under higher-order thinking skills were examined; skills such as creative thinking and problem solving were not addressed. In addition, research can be conducted on the use of AI in the areas of creative and problem-solving thinking skills.

For future research, a similar study can be conducted with a larger sample group using a quantitative method and more generalizable results can be obtained. The participants in this study were graduate students; the results can be compared by applying the study to study groups consisting of students with different educational levels. Studies can be conducted to develop guidelines on how artificial intelligence applications can be used in academic studies by adhering to scientific and ethical principles.

CONCLUSION

In conclusion, in this study, which was conducted to determine whether using CG prompts students to think more deeply through reflection reports, it was observed that most of the participants accepted all the information presented by CG based on a citation as true and did not feel the need to control data reliability, and they could be manipulated by CG while doing self-evaluation. In addition, although there were questions that encouraged them to think critically and reflectively while preparing their reflective reports, in which they compared their essays with CG and although they had taken a graduate level course on the teaching of higher-order thinking skills, it was understood that they could not show the expected performance in using higher order thinking skills except for one participant. It was observed that the reason why two participants did not transfer content from CG's essays to theirs was not due to ethical concerns rather than the data reliability.

REFERENCES

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), 1-13. https://doi.org/10.30935/cedtech/13152

Bisconti, P., Orsitto, D., Fedorczyk, F., Brau, F., Capasso, M., De Marinis, L., & Schettini, C., Eken, H., Bisconti, P., Orsitto, D., Fedorczyk, F., Brau, F., Capasso, M., De Marinis, L., & Schettini, C., Eken, H., Merenda, F., & Forti, M. (2023). Maximizing team synergy in AI-related interdisciplinary groups: an interdisciplinary-by-design iterative methodology. , 38(4), 1443-14. *AI & SOCIETY, 38*(4), 1443-1452. https://doi.org/10.1007/s00146-022-01518-8

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals .* Longman.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). Machine learning: A historical and methodological analysis. *Techniques and Methodology, 4*(3), 69-78.

Chung, H. M., & Silver, M. S. (1992). Rule-based expert systems and linear models: An empirical comparison of learning-by-examples methods. *Decision Sciences, 23*(3), 687-707. https://doi.org/10.1111/j.1540-5915.1992.tb00412.x

Cornish, F., Gillespie, A., & Zittoun, T. (2014). Collaborative analysis of qualitative data. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 79-93). SAGE.

Costa, A. L. (1985). *Developing minds: A resource book for teaching thinking.* Alexandria, VA: Association for Supervision and Curriculum Development.

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

De Angelis, L., Baglivo, F., G., A., G.P., P., P., F., A.E., T., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front. Public Health, 11*, 1-15. https://doi.org/10.3389/fpubh.2023.1166120

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

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? *Frontiers in Education, 8*(May), 1166682.

Hariri, W. (2023). Unlocking the potential of chatgpt: A comprehensive exploration of its applicatıons, advantages, limitations, and future directions in natural language processing. *Computation and Language*, 1-23. https://doi.org/10.48550/arXiv.2304.02017

Iskender, A. (2023). Holy or Unholy? Interview with Open AI's ChatGPT. *European Journal of Tourism Research, 34*, 1-11. https://doi.org/10.54055/ejtr.v34i.3169

Johns, A. M. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly, 20*(2), 247-265. https://doi.org/10.2307/3586543

Kumar, U., Dubey, B., & Kothari, D. P. (2022). *Research methodology: Techniques and trends.* CRC Press.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly, 45*(3), 1501-1525. https://doi.org/10.25300/MISQ/2021/16564

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Digital Health, 5*(3), 105-106. https://doi.org/10.1016/S2589-7500(23)00019-5

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282 .

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). ChatGPT or human? Detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *Computation and Language*, 1-11. https://doi.org/10.48550/arXiv.2301.13852

Plebani, M. (2023). ChatGPT: Angel or Demond? Critical thinking is still needed. *Clinical Chemistry and Laboratory Medicine, 61*(7), 1131–1132. https://doi.org/10.1515/cclm-2023-0387

Rawas, S. (2023). ChatGPT: Empowering lifelong learning in the digital age of higher education. *Education and Information Technologies*, 1-14. https://doi.org/10.1007/s10639-023-12114-8

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sallam, M., Salim, N. A., Barakat, M., & Al-Tammemi, A. B. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive studyhighlighting the advantages and limitations. *Narra J, 3*(1), 1-14. https://doi.org/10.52225/narra.v3i1.103

Schön, D. A. (1992). *The reflective practitioner how professionals think in action.* Routledge.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management, 74*, 1-14. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Swartz, R. J., & Parks, S. (2004). *Infusing the teaching of critical and creative thinking into content instruction.* California: The Critical Thinking Co.

Swartz, S., & McGuinness, C. (2004). *Developing and Assessing Thinking Skills Project Part 1.* Boston: National Center for Teaching Thinking.

Turing, A. M. (2009). Computing machinery and intelligence. In R. R. Epstein (Ed.), *Parsing the Turing Test* (pp. 23-65). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

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

Wegerif, R. (2007). *Literature review in thinking skills, technology and learning.* Open University: Technology and Learning.

Data Availability Declaration

Data Availability Upon Formal Request:

While the primary datasets utilized in this study are not publicly accessible due to certain constraints, they are available to researchers upon a formal request. The authors have emphasized maintaining the integrity of the data and its analytical rigor. To access the datasets or seek further clarifications, kindly reach out to the corresponding author. Our aim is to foster collaborative academic efforts while upholding the highest standards of research integrity.

Author Contributions

All authors, Yalçın Dilekli and Serkan Boyraz contributed equally to this work. They collaboratively handled the conceptualization, methodology design, data acquisition, and analysis. Each author played a significant role in drafting and revising the manuscript, ensuring its intellectual depth and coherence. All authors have thoroughly reviewed, provided critical feedback, and approved the final version of the manuscript. They jointly take responsibility for the accuracy and integrity of the research.

Author(s)' statements on ethics and conflict of interest

Ethics statement: We hereby declare that research/publication ethics and citing principles have been considered in all the stages of the study. We take full responsibility for the content of the paper in case of dispute.

Statement of interest: None.

Funding: None.

Acknowledgements: None.

Biographical notes:

Yalçın Dilekli: He has a Ph.D. degree in the Curriculum and Instruction and his research interests include teaching thinking skills, curriculum needs analysis, qualitative inquiry.

SC Scopus Author Identifier Number: 57190179190

Web of Science Researcher ID: HIZ-8410-2022

Google Scholar Researcher ID: D9YyKAAAAAJ

Serkan Boyraz[3]: He has a Ph.D. degree in the Curriculum and Instruction and his research interests include technology integration in teaching and learning situations, computer assisted teaching and learning, research methodology, quantitative and qualitative data analysis.

SC Scopus Author Identifier Number: 57192646485

Web of Science Researcher ID: GLR-5517-2022

Google Scholar Researcher ID: HNU31fEAAAAJ

---

[3] Corresponding Author