Journal of Pedagogical Research
Volume 9, Issue 4, 2025
https://doi.org/10.33902/JPR.202535435

Research Article

Validity challenge in GenAI models: Evaluating the validity of content generated by text-to-image models in the context of social studies education

Okan Yetişensoy

Bayburt University, Faculty of Education, Türkiye (ORCID: 0000-0002-6517-4840)

Generative artificial intelligence (GenAI) models have led to many positive changes in educational settings; however, the validity of the content they produce remains a significant topic of academic discussion. This research aims to determine the validity of content produced by text-to-image models within the context of social studies education. For this purpose, different curricular outcomes from the social studies curriculum implemented in Türkiye were identified, and image content related to them was generated using DALL-E 3. The validity of this content was assessed within a panel consisting of social studies teachers. Quantitative analyses revealed inconsistencies, showing that some images are sufficient in terms of scientific accuracy as well as socio-cultural and geographical relevance, while others are not. Intraclass correlation coefficient analyses demonstrated that there was significantly moderate agreement among the panelists regarding these evaluations, indicating that the assessments are reliable. Qualitative analysis, on the other hand, revealed that panelists evaluated some images positively in terms of validity, noting that the content produced by these models has significant potential to enhance the learning of relevant outcomes of social studies. However, it was indicated that certain images exhibit prominent algorithmic biases, provide misleading or incorrect information, include details that could be characterized as hallucinations, and could potentially negatively impact the learning process. Additionally, it was determined that some images, in an attempt to highlight a particular cultural element, displays representations that are either disconnected from reality or overemphasized, which could be termed "Socio-cultural algorithmic exaggerations". At this point, it is considered essential for all educators, including social studies teachers, to develop a critical perspective on the content created by GenAI models and to use this content only after thorough evaluation.

Keywords: GenAI; Social studies; Text-to-image models; Validity

Article History: Submitted 15 April 2025; Revised 25 June 2025; Published online 13 August 2025

1. Introduction

Since Alan Turing raised the question "Can machines think?" in his seminal paper "Computing Machinery and Intelligence", published in 1950, artificial intelligence [AI] has seen substantial changes. With the rise of generative models, especially, AI systems have acquired the capacity to produce diverse products that are almost indistinguishable from those created by humans. This advancement has introduced significant shifts in many areas, notably in educational settings (Hsu & Ching, 2023; Zhang et al., 2023). According to many scholars today, generative AI [GenAI] not only represents a growing trend in technology-enhanced learning but also a revolutionary

Address of Corresponding Author

Okan Yetişensoy, PhD, Bayburt University, Faculty of Education, Department of Educational Sciences, 69000, Bayburt, Türkiye.

✉ okanyetisensoy@bayburt.edu.tr

How to cite: Yetişensoy, O. (2025). Validity challenge in GenAI models: Evaluating the validity of content generated by text-to-image models in the context of social studies education. *Journal of Pedagogical Research, 9*(4), 81-101. https://doi.org/10.33902/JPR.202535435

advancement in educational processes (Bozkurt & Sharma, 2024; Hwang & Chen, 2023). While the integration of large language models [LLMs] such as ChatGPT into education is transforming traditional outcome-based practices into process-oriented ones (Chiang et al., 2024), it is also reshaping long-standing pedagogies and fostering a paradigm shift from disciplinary to interdisciplinary learning (Chiu, 2024). This innovation also brings about significant changes in the conventional roles of educational stakeholders (Wang et al., 2024). Teachers, for instance, who once actively participated in every stage of the educational process, are now able to quickly prepare engaging materials and activities that enrich course content, provide prompt assessments and feedback on student work, analyze the necessary data in seconds to gain insights into student needs, and update the teaching process in an efficient manner (Jiang et al., 2024; Yeh, 2024). Students, on the other hand, who once had engaged learning scenarios largely confined within school boundaries, can, by utilizing these models, acquire a learning partner that offers personalized and effective learning experiences irrespective of time and location (Dehghani & Mashhadi, 2024). While the relevant advantages brought by GenAI models are crucial, certain challenges associated with these models also raise significant concerns within the academic community (Kee et al., 2024).

Recently, one of the concepts closely associated with GenAI models stands out as "Validity". Current literature indicates that the term "Validity" is used in several contexts in the use of GenAI in education. One context involves educational assessment processes and examines the validity of automated assessments performed by text-to-text models (Pack et al., 2024) or the various types of questions they generate in a comprehensive framework (Kıyak & Emekli, 2024; Nasution, 2023). The second context, on the other hand, considers validity as a challenge (Suppadungsuk et al., 2023) and focuses on its definition as "The quality of being based on truth or reason or of being able to be accepted" (Dictionary Cambridge, 2024). Studies regarding this context focus on evaluating the scientific accuracy of the content generated by GenAI models and their ability to accurately convey the material they are intended to represent. In such studies, the term validity tends to be used in close relation to the term accuracy (Peled et al., 2024) and this typically includes processes such as expert reviews, comparisons based on document analysis, and evaluations conducted by the authors themselves (Elmas et al., 2024; Nielsen et al., 2023; Zalzal et al., 2023). Even though studies in this context generally focus on the validity of text content created by text-to-text models such as ChatGPT (Wei et al., 2024), there is also an increasing interest in determining the validity of image content produced by text-to-image [T2I] models such as DALL-E, MidJourney, or Stable Diffusion as well. However, studies on this subject is still scarce and there is a significant need for more research specifically aimed at evaluating the validity of content generated by T2I models in education (Fareed et al., 2024).

Social studies stands out as one of the subjects where GenAI models hold significant potential in teaching and learning processes (Berson & Berson, 2023). The existing literature indicates that AI-supported teaching activities implemented in social studies classes show promise for enhancing students' learning performance as well as improving teachers' instructional practices (Yetişensoy & Karaduman, 2024). While the existing literature points to significant GenAI-related challenges that can be addressed in social studies (Yetişensoy, 2025), there is currently no research investigating the validity of GenAI-generated content in the field of social studies education. Given that the use of T2I models in classroom settings is becoming increasingly widespread (Lee et al., 2023; Vartiainen et al., 2023) and the validity issues GenAI models present can lead to significant challenges in educational contexts (Chan & Hu, 2023; Kim et al., 2024), studies addressing the validity of T2I models could provide important perspectives for education stakeholders, including social studies educators. In addition, this kind of study could support social studies teachers in adopting a more critical stance toward AI-generated content and in promoting this awareness among their students. Moreover, such a study is expected to make a significant contribution to the existing literature on both GenAI and technology integration in social studies education. At this point, the next section of the study initially addresses the use of GenAI models in education and

details the validity issue, which is emerging as a significant problem in educational processes. Subsequently, the reflection of validity issues on T2I models is discussed. Finally, the gap in the literature is detailed, and the study's questions are presented.

2. Literature Review

2.1. The Validity Challenge in the Use of GenAI in Education

The rapid expansion of GenAI models in the field of education has introduced numerous new advantages for educational stakeholders. However, this rapid growth has also resulted in significant challenges widely discussed in the academic community (Ogunleye et al., 2024). Some of these challenges include the copyright issues, plagiarism, bias, overreliance on technology, the deepening digital divide, cybersecurity risks, lack of data privacy, accountability and transparency (Chauncey & McKenna, 2023; Dwivedi et al., 2023). Additionally, the issue of GenAI models providing inaccurate, misleading, or fabricated information has been highlighted as a significant concern in many scientific studies (Sallam, 2023; Wittmann, 2023). This issue, often addressed in the literature as the validity challenge (Elmas et al., 2024; Suppadungsuk et al., 2023), fundamentally stem from the fact GenAI models like ChatGPT generate outputs based on highly complex datasets, which sometimes lack factual accuracy and credibility. Moreover, the training data for these models include materials from diverse cultures, which can result in outputs that exhibit specific biases (Yu, 2024). Additionally, the training data these models depend on could be confined to a specific time period. For instance, ChatGPT's knowledge cut-off point is limited to specific dates, and the lack of being up-to-date also plays a significant role in generating inaccurate and misleading information (Hsu & Ching, 2023). Furthermore, the fact that models such as ChatGPT might be unable to access discipline-specific datasets on the internet raises the potential for providing incorrect information in those subject areas (Zarei et al., 2024). Moreover, GenAI models cannot assess the validity of content and determine if the outputs they generate are inaccurate or misleading (Chan & Hu, 2023).

All these shortcomings have raised significant concerns in the field of education (Wang et al., 2024). Montenegro Rueda et al. (2023) emphasized that GenAI models, in educational processes, can produce well-written and seemingly reasonable information that, in reality, has no connection to the truth and can be considered hallucinations. They noted that this situation could negatively impact students' understanding of the topics they are learning. Similarly, Chang and Kidman (2023) pointed out that the information generated by GenAI models could contain significant inaccuracies or biases, and emphasized the growing necessity for students to develop a critical view of the relavant content. Addressing this issue from the perspective of medical education, Lee (2023), stated that medical education requires a high level of precision and accuracy, and even minor misinformation provided by ChatGPT to medical students could lead to significant consequences for patient safety in the future. Bozkurt et al. (2023), on the other hand, emphasized the increasing responsibility of educators in this regard, noting that it is critical for all educators to develop a critical perspective on GenAI-generated information to protect students from misleading content. Furthermore, it is frequently emphasized that these validity issues not only undermine the integrity and credibility of the educational experience (Gill et al., 2024) but also pose a significant obstacle to achieving the goals of various educational curricula (Zarei et al., 2024).

When the literature is examined, it is observed that there is growing interest in studies aimed at examining the validity of content provided by GenAI models, especially text-to-text models, in education. While the current literature indicates that there are studies demonstrating that ChatGPT largely provides accurate answers to user questions (Johnson et al., 2023; Kuşçu et al., 2023; Mackey et al., 2024; Yang & Chang, 2024), a considerable portion of research points to significant shortcomings in terms of validity. This situation is confirmed by many studies in the literature. Elmas et al. (2024), for instance, examined the validity of ChatGPT within the discipline of biochemistry, specifically analyzing its responses to questions about cell energy metabolism through document review. The study found that ChatGPT frequently provided incorrect or

incomplete answers and demonstrated persistence in these incorrect responses. Suppadungsuk et al. (2023) examined the validity of references provided in Vancouver style by ChatGPT and determined that 31% of the 610 references were fabricated and 7% were incomplete. Moreover, it was found that 54% of the references had incorrect DOIs. Chelli et al. (2024) found in a similar study that models like ChatGPT and Bard have a significant rate of hallucination when generating references for systematic reviews, with hallucination rates standing at 39.6% for GPT-3.5, 28.6% for GPT-4, and 91.4% for Bard. Wei et al. (2024) conducted a comprehensive meta-analysis on this subject, analyzing 17 studies in the medical field. They determined that ChatGPT could display an overall integrated accuracy of 56% for medical queries.

In addition to all these, the validity of the content produced by T2I models is emerging as an important point of academic discussion. Chen, Liu et al. (2024) note that current studies evaluating T2I generative models basically focus on two main aspects: image quality and image-text alignment. However, despite the plethora of studies focusing on the quality of T2I-generated images (Bahani et al., 2023; Califano & Spence, 2024) and their alignment with texts (Alikhani et al., 2023; Leivada et al., 2023), there is also a growing interest in assessing the validity of these images as well. In the context of T2I-generated images, validity refers to the degree to which the generated content reflects reality. Accordingly, an image can be regarded as valid if it represents the intended phenomenon in a manner that aligns with both scientific accuracy and socio-cultural relevance (Ajmera et al., 2024; Fared et al., 2024; Sallam, 2023). When the literature is examined, it is seen that various methods are used to evaluate the validity of T2I-generated images. Examples of these evaluation methods include organizing panels and obtaining expert opinions with likert-type validity assessment forms (Buzzaccarini et al., 2024; Seth et al., 2024), using digital applications to detect accuracy of images (Kim & Lee, 2024), or direct evaluations made by the authors themselves (Adams et al., 2023).

The current literature indicates that studies directly evaluating the validity of images created by T2I models are quite limited, and the few that exist are primarily conducted within the fields of medicine and medical education (Lim et al., 2023; Temsah et al., 2024). In a notable study conducted by Ajmera et al. (2024) titled "Validity of ChatGPT-generated musculoskeletal images," musculoskeletal visuals were produced using DALL-E, and the validity of these images was assessed by a panel comprising three radiologists and a musculoskeletal specialist. The study findings indicated significant anatomical errors in the visuals; for instance, many images included bones that do not exist within the skeletal system, incorrect muscle connections, and erroneous joint structures. In a similar study, Seth et al. (2024) generated clinical images of various ulcers using DALL-E, Midjourney, and Blue Willow for surgical education purposes. Subsequently, the validity and educational appropriateness of these images were evaluated by a group of surgeons. The study found that, although these models, especially DALL-E, showed accurate results, they demonstrated significant deficiencies in depicting relevant ulcers with critical details such as depth and colour. Similarly, Kim and Lee (2024) found in their study on landscape design that MidJourney was insufficient in accurately depicting natural elements with detailed characteristics. In a different study, Buzzaccarini et al. (2024) utilized MidJourney to create anatomical images for aesthetic surgery education. These images were then assessed by aesthetic surgery specialists in terms of accuracy, correctness, and visual impact. The findings revealed that all images produced by MidJourney exhibited significant inaccuracies, highlighting their potential to be educationally ineffective and lead to serious misunderstandings

When the relevant studies are generally evaluated, it is observed that the validity of content generated by GenAI models has been examined in many studies, with most focusing on text-to-text models like ChatGPT. Additionally, it is seen that there is growing interest in studies investigating the validity of content generated by T2I models like MidJourney and DALL-E in education. However, these studies are primarily conducted within the context of medical education, and there are no studies addressing this issue in the scope of social studies. At this point, it can be argued that more studies are needed to evaluate the validity of content provided by

T2I models, especially across different disciplines. Moreover, considering that social studies is one of the core subjects in the education systems of many countries, it can be argued that there is also a significant need for studies examining the validity of GenAI-generated content, specifically focusing on T2I models, in the field of social studies education. At this point, it is believed that the relevant study would be beneficial in filling these gaps and contributing to the existing literature on this subject.

2.2. The Present Study

This research aims to determine the validity of content produced by T2I models within the context of social studies education. The research questions to be addressed are as follows:

RQ 1) How do panelist social studies teachers evaluate the validity of image content generated by T2I models for curricular outcomes of social studies from the quantitative perspective?

RQ 2) How do panelist social studies teachers evaluate the validity of image content generated by T2I models for curricular outcomes of social studies from the qualitative perspective?

When the relevant questions are examined, it is observed that the first question of the study aims to evaluate the validity of GenAI-generated image content from a quantitative perspective. The reason for choosing a quantitative method at the initial stage is to minimize researcher-induced subjectivity and to obtain more objective and reliable results supported by statistical data. Additionally, for the second question of the study, a qualitative method was chosen. The reason for this is to enhance and further detail the data obtained from the first question with qualitative findings, thereby presenting a more comprehensive and holistic set of findings.

3. Method

3.1. Research Design

This study was conducted according to the convergent mixed methods design. In this design, the researcher collects the quantitative and qualitative data at roughly the same time and then combines them during the analysis and interpretation stages. This approach provides the researcher with an opportunity to explore research questions in a more comprehensive and detailed manner (Creswell & Creswell, 2018). In this study, the validity of the image content developed for specific social studies outcomes was evaluated through a panel process, where both quantitative and qualitative data were collected. These data were then integrated to present a more comprehensive framework.

3.2. Participants

The participants in this study are five Social Studies teachers, three male and two female, with their teaching experience ranging from 4 to 22 years. The primary criterion for their selection was having a basic understanding of technology and AI integration in education. To assess this, preliminary interviews were conducted, and teachers who expressed confidence in their knowledge of these subjects were included in the study. The decision to focus exclusively on Social Studies teachers, on the other hand, was based on the expectation that, as the primary figures delivering these learning outcomes in a practical context, they would provide more reliable insights.

3.3. Data Collection Tools

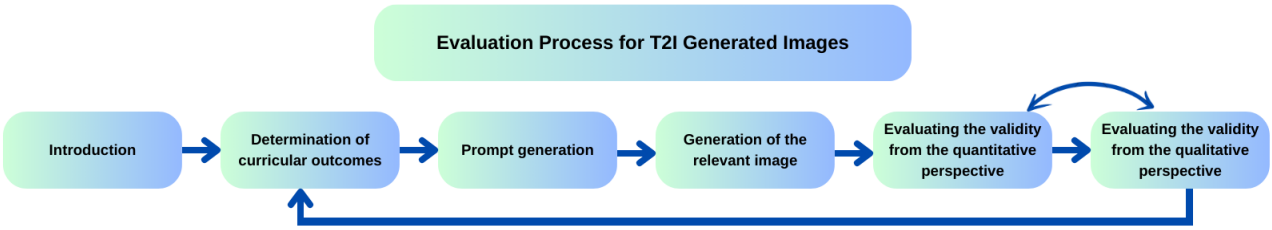
The data for the research was primarily collected using an online validity assessment questionnaire developed by the researcher based on theorical framework regarding the validity concept (Ajmera et al., 2024; Elmas et al., 2024; Nielsen et al., 2023; Peled et al., 2024; Suppadungsuk et al., 2023). In the relevant form, participants were asked to rate each image on a five-point Likert scale based on how well it met four key criteria. The criteria forming the basis of these four sections are the scientific accuracy of the details in the image, freedom from causing misunderstandings and misconceptions, the socio-cultural relevance of the details in the generated

image and the image's general ability to accurately present the intended phenomenon or situation. Additionally, the detailed feedback shared by panelists during the panel was used as a qualitative data source for the research.

3.4. Data Collection Process

The data for this study were systematically collected through an online panel process, moderated by the researcher. The steps underlying this process are presented in Figure 1.

Figure 1
Main steps of the panel process



As seen in Figure 1, the introduction constitutes the first stage of the panel process. During this stage, the researcher delivered a comprehensive presentation on T2I models and the validity challenges associated with them. Following this, information about the purpose and scope of the research was provided, and a brief introduction was held among the panelists. Subsequently, the current Social Studies curriculum (Ministry of National Education [MoNE], 2018) implemented in Türkiye was reviewed with the panelists, and based on their suggestions and mutual agreement, one learning outcome was selected. Afterward, the panelists were asked to relate the selected learning outcome to their teaching experiences and to suggest a prompt for creating an image that could be used in the classroom. A total of three prompt suggestions were received, and the researcher used DALL-E 3 to generate images based on these three prompts. The choice of the relevant T2I model, meanwhile, was influenced by its ability to generate complex outputs even from simple sentences (OpenAI, 2023), as well as its user-friendly interface that does not require integration with external platforms.

Immediately after all the images were generated, the panelists were asked to review the images in sequence and then rate them using the validity assessment forms. Following the evaluation, detailed feedback on each image was requested from the panelists, and the validity of the images was discussed among them. All processes, such as reviewing the curriculum for choosing curricular outcome, generating images, and presenting them to the panelists, were managed through the moderator's screen. Additionally, panelists were allowed to update their evaluations on the validity assessment form throughout the panel process. A total of four different learning outcomes were reviewed, and this cycle was repeated for each outcome starting from the "Determination of curricular outcome" stage. While the panel concluded after two sessions, the curricular outcomes examined during the process are as follows (MoNe, 2018):

- 4th-grade curricular outcome: "Students prepare for natural disasters".
- 5th-grade curricular outcome: "Students recognize the significant contributions of Anatolian and Mesopotamian civilizations to human history based on tangible artifacts".
- 6th-grade curricular outcome: "Students examine Türkiye's main physical geography features, including landforms, climate characteristics, and vegetation, on specific maps".
- 7th-grade curricular outcome: "Students plan their career choices by considering new professions that have emerged due to global developments".

3.5. Data Analysis and Credibility

In the analysis of the study's quantitative data, descriptive statistics were primarily used. At this point, each image's score on the five-point likert scale was shown alongside the criteria to offer an overall perspective on the validity level of the content. Additionally, to determine the reliability of the panelists' evaluations, intraclass correlation coefficient [ICC] calculations were performed using Statistical Package for the Social Sciences [SPSS] 29. The ICC value calculations conducted on a two-way mixed-effects model with the type of absolute agreement suggested that the panelists' ratings for each learning outcome demonstrated a moderate level of reliability and were statistically significant. In this context, all evaluations were included in the analyses (Koo & Li, 2016). For the analysis of the study's qualitative data, deductive content analysis was employed. This type of analysis is recommended when the structure of the analysis is established on the basis of previous knowledge (Elo & Kyngäs, 2008). In this context, based on existing literature on the concept of validity (Ajmera et al., 2024; Elmas et al., 2024; Nielsen et al., 2023; Peled et al., 2024; Suppadungsuk et al., 2023), "The validity challenge regarding T2I models" was identified as the main category/theme, while "Scientific accuracy of the details", "Freedom from causing misunderstandings and misconceptions" and "Socio-cultural and geographical relevance" were identified as generic categories. The key concepts/codes mentioned in the panelist statements were summarized under these generic categories, resulting in a structure that moves from general to specific (see Appendix 1). Additionally, to ensure the credibility of the qualitative data, member checking and peer debriefing (Lincoln & Guba, 1985) were utilized. As part of peer debriefing, a social studies educator with expertise in qualitative research was involved in the qualitative analysis process. The researcher and the peer held two consensus meetings to reach an agreement on the analyses. Additionally, as part of member checking, the qualitative data from the panel process were transcribed from the recordings and subsequently shared with the participants to confirm the accuracy of their statements.

4. Findings

4.1. Findings regarding the Images Developed for the First Learning Outcome

The first image content in the scope of the study was developed for the 4th-grade curricular outcome entitled "Students prepare for natural disasters". The relevant content is presented in Figure 2.

Figure 2

Image content regarding the first curricular outcome

Promt: Could you generate an image showing a Turkish child preparing an earthquake emergency bag in a home setting	**Prompt:** Could you generate an illustration of an earthquake drill in a Turkish school classroom?	**Prompt:** Could you provide me an illustration of a Turkish family preparing their home for an earthquake by securing potentially toppling items.

The ratings that the panelists provided regarding the relevant images, based on four key criteria, are presented in Table 1.

Table 1
The panelists' ratings for the images generated for the first learning outcome

Criteria	*Image Number*	*P1*	*P2*	*P3*	*P4*	*P5*	*Overall*
The scientific accuracy of the details in the generated image	1A	5	3	5	3	5	4.2
	1B	2	2	3	2	2	2.2
	1C	4	4	5	3	5	4.2
The image's freedom from causing misunderstandings and misconceptions.	1A	5	4	5	4	4	4.4
	1B	2	1	2	2	2	1.8
	1C	4	5	5	3	5	4.4
The socio-cultural and geographical relevance of the details in the generated image	1A	3	3	4	4	4	3.6
	1B	3	3	5	5	3	3.8
	1C	4	5	5	5	5	4.8
The image's general ability to accurately present the intended phenomenon or situation.	1A	5	3	5	3	5	4.2
	1B	2	3	2	2	3	2.4
	1C	4	4	4	3	5	4
Intraclass Correlation Coefficient Value: .527, p = .006							

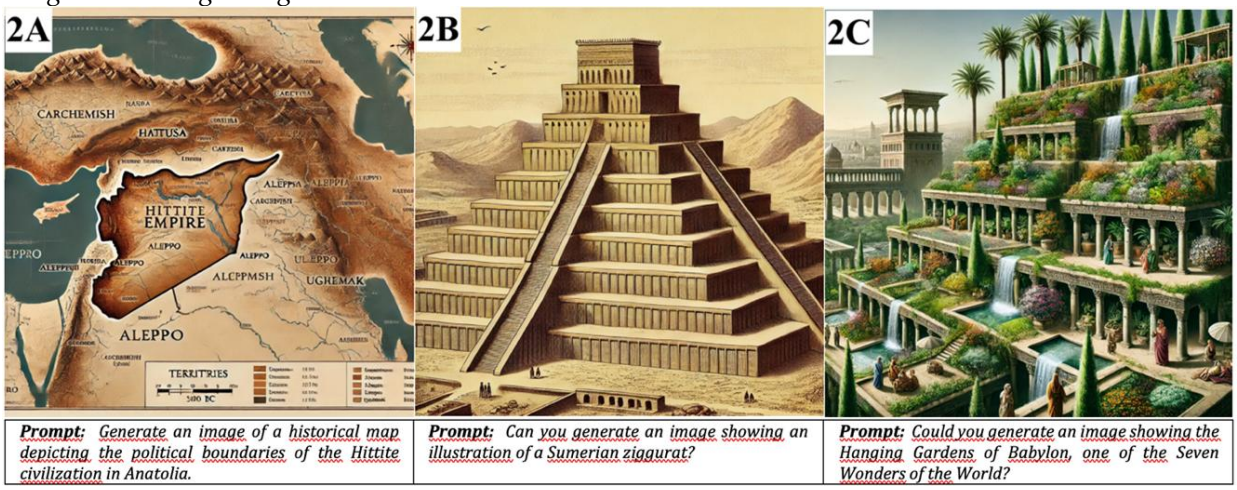
As seen in Table 1, the panelists predominantly rated images 1A and 1C with scores of 4 and 5, indicating that these images, with their current validity, largely met the relevant criteria. However, image 1B showed significant insufficiencies in this regard. The ratings provided by the panelists revealed that, aside from the socio-cultural and geographical context criterion, the image's ability to meet the other three criteria was below average. During the panel process, panelists provided insights that detailed the quantitative results. They particularly noted that the details in images 1A and 1C were accurate from a scientific standpoint. The depiction of items such as a flashlight, water, and medical supplies in the emergency kit in image 1A, as well as the reinforcement of the television in image 1C for earthquake preparedness, were positively received in terms of supporting learning the curricular outcomes. However, P1 and P4 raised doubts that DALL-E may have overlooked subtle nuances, noting that a typical emergency kit and its contents, such as water, a flashlight, and medical supplies, should be compact and easy to carry. Additionally, they suggested that securing furniture like cabinets, which are more likely to tip over, would be more appropriate than focusing on the television. Moreover, all the panelists expressed concerns that the portrayal of the earthquake drill in image 1B could cause significant misunderstandings among students regarding proper drill procedures. P4, for instance, summarized their thoughts with the following statement: "This image looks really chaotic. The students are all in different, mostly wrong positions, and the teacher hasn't taken any position either. It really shows that the drill is being done completely wrong".

In addition, all the images were found to be mostly successful by the panelists in reflecting the socio-cultural and geographical context in which the events took place. For instance, P2 stated that relevant images effectively captured the context of Turkish culture through the use of furniture, flags, and decorations, noting that the details in image 1C were particularly exceptional in this regard. However, P1 and P3 pointed out that the text in the images was not in Turkish, highlighting a significant linguistic bias. Similarly, P4 expressed that while health products in Türkiye typically feature the Red Crescent emblem, the products in image 1A displayed a white cross, indicating the presence of cultural bias. They also noted that these circumstances could lead to the misrepresentation of culture. The panelists also indicated some exaggerations in the depiction of the socio-cultural context. For example, P2 and P3 mentioned that image 1B depicted a Turkish carpet covering the entire classroom, which they found quite bizarre and unrealistic since carpets are only used in preschool classrooms in the Turkish education system. Similarly, P1 stated that the Turkish motifs in image 1C were exaggerated and that combining these traditional and rarely used motifs with modern home design created a peculiar image.

4.2. Findings regarding the Images Developed for the Second Learning Outcome

The second image content within the scope of the study was developed for the 5th-grade curricular outcome entitled "Students recognize the significant contributions of Anatolian and Mesopotamian civilizations to human history based on tangible artifacts". The relevant content is presented in Figure 3.

Figure 3
Image content regarding the second curricular outcome



The ratings that the panelists provided regarding the relevant images, based on four key criteria, are presented in Table 2.

Table 2
The panelists' ratings for the images generated for the second learning outcome

Criteria	*Image Number*	*P1*	*P2*	*P3*	*P4*	*P5*	*Overall*
The scientific accuracy of the details in the generated image	2A	1	2	2	2	2	1.8
	2B	4	4	5	5	5	4.6
	2C	3	4	5	4	5	4.2
The image's freedom from causing misunderstandings and misconceptions.	2A	1	2	2	1	2	1.6
	2B	4	4	5	5	4	4.4
	2C	3	5	5	4	5	4.4
The socio-cultural and geographical relevance of the details in the generated image	2A	2	3	3	3	4	3
	2B	3	3	5	5	5	4.2
	2C	4	4	5	4	5	4.4
The image's general ability to accurately present the intended phenomenon or situation.	2A	1	2	2	2	2	1.8
	2B	3	3	5	5	5	4.2
	2C	4	4	5	5	5	4.6
Intraclass Correlation Coefficient Value: .692, $p < .001$							

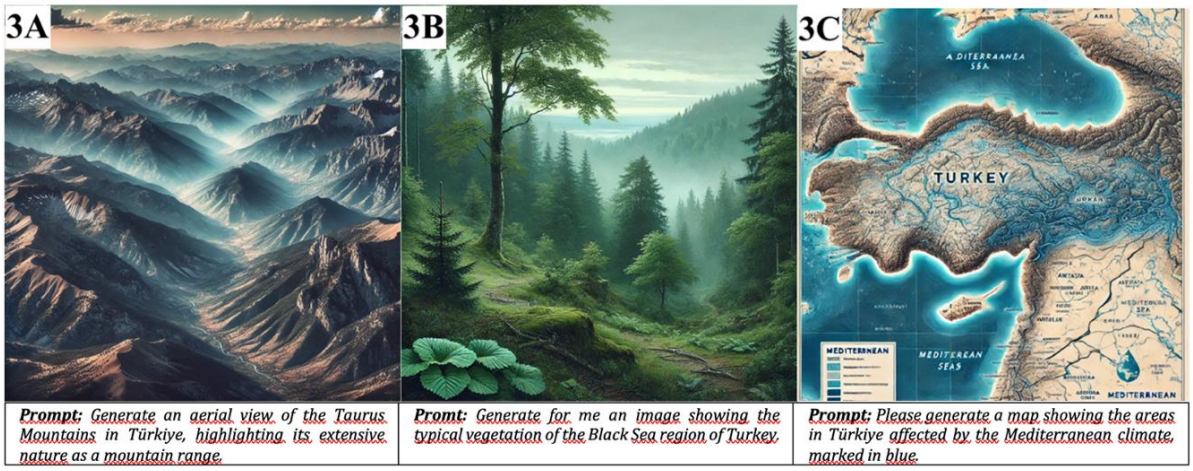
As seen in Table 2, the panelists tended to rate the images 2B and 2C with scores of 4 and 5 across all criteria, indicating that these images demonstrated significant validity. In contrast, the scores for image 2A were mostly below average. During the panel process, the panelists noted that the images 2B and 2C largely demonstrated scientific accuracy, while 2A was found to be significantly lacking in this regard. The panelists emphasized that the Hittite state was located in Anatolia, but in the image, it was shown in the Middle Eastern region, which could lead to misunderstandings among students. Additionally, the presence of meaningless place names written in an unclear alphabet was also criticized by P2 and P5. P1, for instance, described these fabricated place names, which do not actually exist, as a case of hallucination and stated that such instances could lead to a decrease in trust toward T2I models within educational processes.

Although the panelists found the depiction of the ziggurat in 2B to be largely accurate, they pointed out that the image content was incomplete in certain aspects. P2 summarized their thoughts on this matter as follows. "We know ziggurats had multiple uses. The lower levels were for storing supplies, the middle ones had temples and libraries, and the top levels were observatories. But none of that is shown in the image". In addition to all these, image 2C was considered the most successful by the panelists. P1 and P2 noted that the image effectively captured the details described in stories about the Hanging Gardens of Babylon, and that the depiction of the gardens in an arid setting was both appropriate and consistent with the geographical context.

4.3. Findings regarding the Images Developed for the Third Learning Outcome

The third image content within the scope of the study was developed for the 6th-grade curricular outcome entitled "Students examine Türkiye's main physical geography features, including landforms, climate characteristics, and vegetation, on specific maps". The relevant content is presented in Figure 4.

Figure 4
Image content regarding the third curricular outcome



The panelists' ratings of the images presented in Figure 4 are provided in Table 3.

Table 3
The panelists' ratings for the images generated for the third learning outcome

Criteria	*Image Number*	*P1*	*P2*	*P3*	*P4*	*P5*	*Overall*
The scientific accuracy of the details in the generated image	3A	4	4	5	4	5	4.4
	3B	4	5	4	4	5	4.4
	3C	1	2	1	2	2	1.6
The image's freedom from causing misunderstandings and misconceptions.	3A	4	4	3	5	5	4.2
	3B	4	5	4	4	5	4.4
	3C	1	2	1	1	2	1.4
The socio-cultural and geographical relevance of the details in the generated image	3A	3	4	3	3	5	3.6
	3B	4	4	4	5	5	4.4
	3C	2	2	3	1	3	2.2
The image's general ability to accurately present the intended phenomenon or situation.	3A	4	4	4	5	5	4.4
	3B	4	5	4	5	4	4.4
	3C	1	2	1	1	2	1.4
Intraclass Correlation Coefficient Value: .509; $p < .001$							

As indicated in Table 3, the panelists tended to rate images 3A and 3B with scores of 4 and 5, demonstrating that these images largely met the relevant criteria. However, image 3C was found

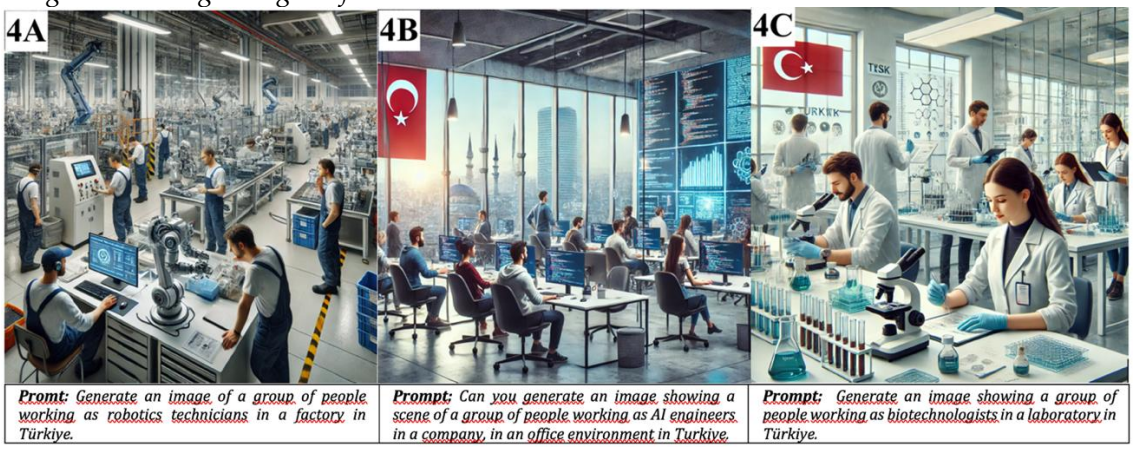
to be insufficient in meeting the criteria across all aspects. In the panel process, panelists stated that images 3A and 3B, developed for the relevant learning outcome, scientifically accurate. For instance, P4 summarized their thoughts on image 3A as follows: "The Taurus Mountains look almost real here, and the image does a great job showing the structure of the mountain range. The small river flowing through the valley looks really realistic too". P2, who evaluated image 3B, noted that while it didn't clearly depict the diversity of trees in the region, the image effectively represented the green nature of the Black Sea climate and could help students grasp its general characteristics. On the other hand, no details were identified in the relevant images that could lead to misunderstandings or conceptual confusion among the students. However, all panelists found image 3C to be quite inadequate in this regard. For instance, P3 and P5 pointed out that the areas shown in blue were actually under the influence of a continental climate, rather than a Mediterranean one. P4 emphasized that such a map could lead to significant misunderstandings among students. Furthermore, P5 pointed out that this image also contains fabricated place names that do not actually exist. P1, on the other hand, described certain elements in the image as quite misleading, such as the Black Sea being labeled with a fabricated and unheard-of name, and a river in the Middle East being referred to as the Mediterranean Sea.

4.4. Findings regarding the images developed for the fourth learning outcome

The fourth image content within the scope of the study was developed for the 7th-grade curricular outcome entitled "Students plan their career choices by considering new professions that have emerged due to global developments". The relevant content is presented in Figure 5.

Figure 5

Image content regarding the fourth curricular outcome



The panelists' ratings of the relevant images based on four key criteria are presented in Table 4. As seen in Table 4, the panelists predominantly rated all the images with scores of 4 and 5 across the four criteria. This result indicates that all the images created for the relevant learning outcome were found to be quite accurate and free of issues, unlike the previous ones. Additionally, the panelists stated that the three images accurately depicted the related professions and could help students conceptualize the professions without leading to misunderstandings. P2, who evaluated how well the images reflected the socio-cultural context, noted that image 4B was the most exceptional in this regard, 4C was also successful, but 4A was weaker. Similarly, P4 praised the adequacy of 4B in reflecting the socio-cultural context but noted that the issue of exaggeration seen in some visuals developed for previous learning outcomes was also present here. P4 particularly highlighted the depiction of a mosque being nearly the same height as a skyscraper in the background as an exaggerated portrayal. In addition to all these, P3 and P5 emphasized that after observing how accurate the images developed for this learning outcome reflected the content, they

Table 4
The panelists' ratings for the images generated for the fourth learning outcome

Criteria	*Image Number*	*P1*	*P2*	*P3*	*P4*	*P5*	*Overall*
The scientific accuracy of the details in the generated image	4A	5	5	5	4	5	4.8
	4B	4	5	5	4	5	4.6
	4C	4	4	5	5	5	4.6
The image's freedom from causing misunderstandings and misconceptions.	4A	5	5	5	4	5	4.8
	4B	4	5	5	5	4	4.6
	4C	4	5	5	5	5	4.8
The socio-cultural and geographical relevance of the details in the generated image	4A	3	2	3	3	4	3
	4B	4	5	5	5	4	4.6
	4C	3	5	3	5	5	4.2
The image's general ability to accurately present the intended phenomenon or situation.	4A	4	4	5	4	5	4.4
	4B	3	5	5	5	5	4.6
	4C	4	5	5	5	5	4.8
Intraclass Correlation Coefficient Value: . 663, p = .006							

got the impression that while DALL-E is quite successful in creating general depictions, it falls short when it comes to accurately representing subjects that require technical knowledge.

5. Discussion

In this study, image content generated by GenAI for various social studies curricular outcomes was assessed by a panel of social studies teachers. The findings revealed discrepancies concerning the validity of these images. The quantitative data indicated that the panelists found a portion of the images to be sufficient in terms of scientific accuracy and in avoiding misinformation and misconceptions. The qualitative data, derived from the panelists' statements, further supported this finding, suggesting that these images, with their current validity, hold significant educational potential. Even though research on the integration of T2I models into educational processes is quite limited (Vartiainen & Tedre, 2023), the few practical studies suggest that these models show promise for enhancing teaching and learning processes (Chen, Zhang et al., 2024; Lee et al., 2023; Vartiainen et al., 2023). Moreover, several theoretical studies highlight the considerable potential of these models in social studies (Karaduman & Yetişensoy, 2023; Yeşilyurt et al., 2025). Considering that the effectiveness of these models primarily depends on the quality of the training datasets, which should include high-quality and detailed images relevant to a specific subject area (Koga & Du, 2024), it can be argued that enriching the datasets used by the relevant T2I models to be more sensitive to different pedagogical areas could enable them to generate more valid content. It can be thought that such image content, with a high degree of validity, not only helps prevent misconceptions during learning but also contributes to more meaningful and engaging educational experiences for both students and teachers across social studies and other subject areas.

In this study, the quantitative data revealed that some of the images were insufficient in terms of scientific accuracy and in avoiding misunderstandings or misconceptions. The qualitative findings suggested that these issues were more likely to arise in content requiring technical knowledge. For example, a developed historical map was found to demonstrate the location of the relevant state in a completely different area, while the geographical map incorrectly displayed the regions where a certain climate was supposed to occur, along with numerous incorrect place names. Similarly, the positions shown in the earthquake drill image, where people protect themselves, were found to be technically inappropriate by the panelists. The current literature suggests that T2I models often fall short in generating accurate content on subjects that require technical knowledge (Temsah et al., 2024) as well as in depicting abstract concepts (Liao et al., 2024). In a study that indirectly addresses this topic, Fareed et al. (2024) organized two workshops involving educators and students from various age groups. In these workshops, students generated images of architectural structures from different historical periods using T2I models.

The results indicated significant architectural inaccuracies in the generated images, such as those of Mesopotamian ziggurats. In this regard, it can be considered that having the content generated by these models, especially in areas requiring technical knowledge, reviewed by specific expert groups and conducting fine-tuning on areas where shortcomings are identified could help reduce inaccuracies and enhance the models' ability to produce valid content on technical subjects. Moreover, such efforts may contribute to addressing the hallucination problem frequently observed in T2I models, as also reflected in the findings of this study. Indeed, the issue of hallucination in T2I models emerges as a significant challenge, particularly in educational contexts (Lim et al., 2025), and it can be argued that this problem negatively affects the trust placed in these models within educational settings. Beside this, considering that adults have a better capacity to recognize inaccuracies in GenAI-generated content, while children, as one of the most vulnerable groups in society, are less equipped to do so (Kruse et al., 2023), raising awareness among young students about the potential for these models to cause misconceptions and inaccurate learning, and encouraging them to develop a critical perspective towards this content, can be regarded as an essential requirement. Additionally, teachers are among the primary stakeholders in education. As practitioners of these models in classroom settings, they require an advanced understanding of AI (Sanusi et al., 2022). At this point, it can also be regarded essential for all educators to develop a critical perspective on the content generated by GenAI models and to use this content only after thorough evaluation.

Another point on which the panelists provided detailed evaluations was the socio-cultural and geographical relevance of the details in the generated images. The panelists emphasized that DALL-E generally showed exceptional performance in this context through details such as architecture and decoration. However, they also noted that certain images exhibited shortcomings that could be interpreted as algorithmic biases in linguistic and cultural aspects. When the literature is examined, it is seen that the biases reflected in the content generated by T2I models stand out as a significant issue (Friedrich et al., 2024; Wang et al., 2023). In a study conducted by Adams et al. (2023) it was determined that X-Ray images produced by DALL-E 2 tend to represent male bodies and exhibit significant gender bias. Similarly, a detailed study on this subject revealed that the Stable Diffusion's depictions of Australian and New Zealander individuals tend to ignore indigenous groups and instead represent these countries' people as fair-skinned European immigrants. Furthermore, the same study observed a tendency to portray Latin American women in a more sexualized manner, while depictions of women from the United Kingdom appeared more conventional (Ghosh & Çalışkan, 2023). In addition, different studies in the literature indicated that these models exhibited significant occupational biases in the images they generated for certain professions (Cheong et al., 2024) and portrayed people with disabilities using distressing details such as decaying skin, sunken eyes, shadow-covered faces, and tattered or dirty clothes (Mack et al., 2024). At this point, it can be argued that the use of T2I models in educational settings may also potentially have unsettling effects on students from specific groups. The current literature indicates that the content generated by T2I models may lead to psychological discomfort in individuals (Xu et al., 2024). Therefore, raising awareness among students about the potential for these models to produce biased content, which may lead to discomfort, can be seen a necessary consideration. A comprehensive study on this subject was conducted by Vartiainen et al. (2024) with the participation of elementary and secondary school students in Finland, where students engaged in a series of hands-on activities using Midjourney to explore algorithmic biases generated by T2I models. The study results found that there were significant improvements in students' existing understanding of algorithmic bias. At this point, it can be considered that conducting similar studies would contribute to enhancing awareness of the subject and encouraging further research in this area.

An additional point the panelists indicated in this study that some images depicted overemphasized or unrealistic representations while conveying the socio-cultural context of events. For instance, when the model was asked to generate an image of an earthquake drill in a

Turkish classroom, it generated an image showing a traditional Turkish rug covering the entire floor of a modern classroom. This detail was considered quite bizarre and unrealistic by the panelists. Similarly, when the model was asked to generate an image of a Turkish family securing their belongings against an earthquake, it combined very traditional and rarely used decorations with a modern home design. This combination was also perceived by them quite odd. Although the scarcity of studies on this topic limits relevant comparisons, in a similar study by Ghosh et al. (2024), participants noted that images related to Indian culture, generated by Stable Diffusion, contained significant exaggerations in their colorful depictions. Magomere et al. (2025), on the other hand, found that T2I models tended to use exaggerated representations when producing images of traditional dishes from different countries' cultural heritage. While there are studies pointing to the potential of GenAI algorithms to encourage the production of exaggerated content (Nguyen et al., 2024; Zack et al., 2024), a specific term describing this particular phenomenon has not yet been identified in the existing literature. Therefore, this study introduces the concept of "Socio-cultural algorithmic exaggeration" to explain the phenomenon and defines this concept as the tendency of algorithms to present representations that are either disconnected from reality or overly emphasized, with the intention of highlighting a particular socio-cultural element. Zhang et al. (2024), who emphasized that T2I models have significant deficiencies in appropriately representing cultural aspects, pointed out that this challenge could jeopardize the accurate expression of cultures and increase the risk of reinforcing cultural misconceptions. In this context, it can be suggested that to mitigate the relevant risks and enable these models to more accurately reflect cultural elements in image content, it would be beneficial to include images with richer depictions of these elements in the datasets, conduct fine-tuning on the models, establish mechanisms for continuous feedback from both experts and general users, and further improve the models' internal mechanisms.

In addition to all these points, it can be argued that the validity issues identified throughout the study may also be attributed to prompt quality to a minor extent. When the prompts provided by the panelists are examined, it is observed that they were generally superficial in nature. Given that the quality of a prompt plays a significant role in shaping the outputs produced by GenAI models (Knoth et al., 2024) it can be assumed that the quality of the prompts in this study also had a certain level of influence on the level of validity. However, considering the presence of hallucinations such as non-existent place names, technical inaccuracies particularly evident in complex structures like maps, and algorithmic biases that are inconsistent with the cultural context, it can be thought that these validity-related issues may be attributed not only to prompt quality, but also, to a significant degree, to systemic deficiencies inherent to the model itself. Nonetheless, the swift progress of these models in recent years offers considerable promise in overcoming the systemic limitations associated with T2I models. Indeed, considering that until the second quarter of 2023, T2I models were almost incapable of accurately depicting human hands and fingers, but they have now advanced to a level where they can realistically portray even highly complex and intricate scenes (Morton, 2024), it can be expected that validity issues in these models will decrease in the near future. It can also be expected that this progress will provide a significant opportunity for these models to generate image content that is not only consistent with scientific and socio-cultural contexts but also of higher educational quality.

6. Limitations and Future Research

This study has certain limitations. First of all, conducting the study with five panelists, all Social Studies teachers, constitutes a significant limitation, particularly in terms of generalizing the quantitative findings. However, considering that the study was conducted using a convergent mixed methods design incorporating a qualitative paradigm, and that qualitative research tends to interpret each situation within its own dynamics without a focus on generalization, it is believed that, despite the lack of generalizability of the quantitative findings, the study provides essential insights, particularly through its qualitative results. Additionally, since examining images within a

panel process is time-consuming, the study was conducted by examining 12 images based on 4 curricular outcomes from the Social Studies curriculum. Although this study is expected to provide an insightful perspective on the concept of validity in educational processes, it is believed that a study examining a larger number of images would provide a much more detailed findings. At this point, it is suggested that studies be conducted to examine the validity of a much larger number of images in educational contexts. Moreover, the exclusive use of a single model, namely DALL·E 3, throughout the study without considering alternative models can be viewed as another limitation. In this regard, conducting similar studies that also incorporate models such as Stable Diffusion or MidJourney in addition to DALL-E is recommended, as such efforts may lead to more comparable results.

Furthermore, the fact that studies examining content produced by T2I models are predominantly conducted within the context of medicine and medical education also limited potential comparisons in the discussion section of this study. At this point, it can be seen as a significant necessity to conduct similar studies in different pedagogical areas such as science, mathematics, history, and geography, as well as social studies education. In addition, the lack of a general consensus on the concept of validity stands out as a factor that could potentially restrict the scope of the researchs on this subject. The current literature indicates that there is no clear consensus on the meaning of "Validity" for content generated by T2I GenAI models. While most studies consider it within the scope of the scientific accuracy of the image content, like this study, others include aspects such as the educational appropriateness or quality of the images as well. As a result, determining the criteria forming the basis of findings of this study posed a significant challenge. Ultimately, validity was decided to be assessed around the scientific accuracy of the details in the image content, its freedom from causing misunderstandings and misconceptions, its socio-cultural and geographical relevance, its general ability to accurately depict the intended phenomenon. The findings regarding the pedagogical value of the images, on the other hand, was presented through the indirect expressions used by the participants in the panel process. Although this study argues that the concept of validity for T2I models should be confined to the accuracy level of the images, it is recommended that further studies be conducted to define the boundaries of "Validity" for both text-to-text and T2I models and to establish a generally acceptable consensus on this matter.

Declaration of interest: The author declares that no competing interests exist.

Data availability: The availability of data can be obtained by requesting directly to the corresponding author.

Ethical declaration: Throughout the study, all ethical guidelines were followed, and informed consent was obtained from all participants.

Funding: No funding was reported for this study.

References

Adams, L.C., Busch, F., Truhn, D., Makowski, M.R., Aerts, H.J.W.L., & Bressem, K.K. (2023). What does DALL-E 2 know about radiology? *Journal of Medical Internet Research*, *25*, 43110. https://doi.org/10.2196/43110

Ajmera, P., Nischal, N., Ariyaratne, S. Botchu, B., Bhamidipaty, K.D.P., Iyengar, K.P., Ajmera, S.R., Jenko, N. & Botchu, R. (2024). Validity of ChatGPT-generated musculoskeletal images. *Skeletal Radiology*, *53*, 1583-1593. https://doi.org/10.1007/s00256-024-04638-y

Alikhani, M., Khalid, B., & Stone, M. (2023). Image-text coherence and its implications for multimodal AI. *Frontiers in Artificial Intelligence 6*, 1048874. https://doi.org/10.3389/frai.2023.1048874

Bahani, M., El Ouaazizi, A., Maalmi, K. (2023). The effectiveness of T5, GPT-2, and BERT on text-to-image generation task. *Pattern Recognition Letters*, *173*, 57-63. https://doi.org/10.1016/j.patrec.2023.08.001

Berson, I. R. & Berson, M. J. (2023). The democratization of AI and its transformative potential in social studies education. *Social Education*, *87*(2), 114-118.

Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C., Romero-Hall, E., Koutropoulos, A., Toquero, C. M., Singh, L., Tlili, A., Lee, K., Nichols, M., Ossiannilsson, E., Brown, M., Irvine, V., Raffaghelli, J. E., Santos-Hermosa, G., Farrell, O., Adam, T., Thong, Y. L., Sani-Bozkurt, S., Sharma, R. C., Hrastinski, S., & Jandrić, P. (2023). Speculative futures on ChatGPT and Generative artificial intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, *18*(1), 53-130. https://doi.org/10.5281/zenodo.7636568

Bozkurt, A., & Sharma, R. C. (2024). Are we facing an algorithmic renaissance or apocalypse? Generative AI, ChatBots, and emerging human-machine interaction in the educational landscape. *Asian Journal of Distance Education*, *19*(1), 1-12. https://doi.org/10.5281/zenodo.10791959

Buzzaccarini, G., Degliuomini, R.S., Borin, M., Fidanza, A., Salmeri, N., Schiraldi, L., Di Summa, P.G., Vercesi, F., Vanni, V.S., Candiani, M., & Pagliardini, L. (2024). The promise and pitfalls of AI-generated anatomical images: Evaluating Midjourney for aesthetic surgery applications. *Aesthetic Plastic Surgery, 48*, 1874-1883. https://doi.org/10.1007/s00266-023-03826-w

Califano, G., & Spence, C. (2024). Assessing the visual appeal of real/AI-generated food images. *Food Quality and Preference*, *116*, 105149. https://doi.org/10.1016/j.foodqual.2024.

Chan, C.K.Y., & Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(43), 1-18. https://doi.org/10.1186/s41239-023-00411-8

Chang, C.H., & Kidman, G. (2023). The rise of generative artificial intelligence (AI) language models-challenges and opportunities for geographical and environmental education. *International Research in Geographical and Environmental Education*, *32*(2), 85-89. https://doi.org/10.1080/10382046.2023.2194036

Chauncey, S.C., & McKenna, H.P. (2023). A framework and exemplars for ethical and responsible use of AI Chatbot technology to support teaching and learning. *Computers and Education: Artificial Intelligence, 5*, 100182. https://doi.org/10.1016/j.caeai.2023.100182

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.L., Clowez, G., Boileau, P., & Ruetsch Chelli, C., (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, *22*(26), e53164. https://doi.org/10.2196/53164

Chen, M., Liu, Y., Yi, J., Xu, C., Lai, Q., Wang, H., Ho, T., Xu, Q. (2024). *Evaluating text-to-image generative models: An empirical study on human image synthesis* (Publication no: 2403.05125). Arxiv. https://arxiv.org/pdf/2403.05125

Chen, Y., Zhang, X., & Hu, L. (2024). A progressive prompt-based image-generative AI approach to promoting students' achievement and perceptions in learning ancient Chinese poetry. *Educational Technology & Society*, *27*(2), 284-305. https://doi.org/10.30191/ETS.202404_27(2).TP01

Cheong, M., Abedin, E., Ferreira, M., Reimann, R., Chalson, S., Robinson, P., Byrne, J., Ruppanner, L., Alfano, M., & Klein, C. (2024). Investigating gender and racial biases in DALL-E mini images. *ACM Journal on Responsible Computing, 1*(2), 1-20. https://doi.org/10.1145/3649883

Chiang, Y. V., Chang, M., & Chen, N.S. (2024). Can generative AI help realize the shift from an outcome-oriented to a process-outcome-balanced educational practice? *Educational Technology & Society*, *27*(2), 347-385. https://doi.org/10.30191/ETS.202404_27(2).TP04

Chiu, T.K.F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence*, *6*, 100197. https://doi.org/10.1016/j.caeai.2023.100197

Creswell, J.H., & Creswell, J.D. (2018). *Research design: Qualitative, quantitative and mixed methods approaches*. Sage.

Dehghani, H., & Mashhadi, A. (2024). A. Exploring Iranian english as a foreign language teachers' acceptance of ChatGPT in english language teaching: Extending the technology acceptance model. *Education and Information Technologies*, *29*, 19813-19834. https://doi.org/10.1007/s10639-024-12660-9

Dictionary Cambridge (2024). *Validity*. Author. https://dictionary.cambridge.org

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Wright, R., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., & Buhalis, D., Carter, L., & 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

Elmas, R., Adiguzel-Ulutas, M. & Yılmaz, M. (2024). Examining ChatGPT's validity as a source for scientific inquiry and its misconceptions regarding cell energy metabolism. *Education and Information Technologies*, *29*, 25427-25456. https://doi.org/10.1007/s10639-024-12749-1

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107-115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Fareed, M.W., Bou, Nassif, A., & Nofal, E. (2024). Exploring the potentials of artificial intelligence image generators for educating the history of architecture. *Heritage*, *7*, 1727-1753. https://doi.org/10.3390/heritage7030081

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., & Kersting, K. (2024). Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, *5*, 2103-2123. https://doi.org/10.1007/s43681-024-00531-5

Ghosh, S. Venkit, P.N., Gautam, S., Wilson, S., & Caliskan, A. (2024). *Do generative AI models output harm while representing non-western cultures: Evidence from a community-centered approach* (Publication no: 2407.14779). Arxiv. https://arxiv.org/abs/2407.14779

Ghosh, S., & Caliskan, A. (2023). Person' == Light-skinned, western man, and sexualization of women of color: Stereotypes in Stable Diffusion. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6971-6985). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.465

Gill, S.S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A.K., Stankovski, V., Abraham, A., Ghosh, S.K., Lutfiyya, H., Kanhere, S.S., Bahsoon, R., Rana, O., Dustdar, S, Sakellariou, R., Uhlig, S., & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber-Physical Systems*, *4*, 19-23. https://doi.org/10.1016/j.iotcps.2023.06.002

Hsu, Y.C., & Ching, Y.H.(2023). Generative artificial intelligence in education, part one: The dynamic frontier. *TechTrends, 67*, 603-607. https://doi.org/10.1007/s11528-023-00863-9

Hwang, G.J., & Chen, N.S. (2023). Editorial position paper: Exploring the potential of generative artificial intelligence in education: Applications, challenges, and future research directions. *Educational Technology & Society*, *26*(2), 1-18. https://doi.org/10.30191/ETS.202304_26(2).0014

Jiang, Y., Xie, L., Lin, G., & Mo, F. (2024). Widen the debate: What is the academic community's perception on ChatGPT?. *Education and Information Technologies, 29,* 20181-20200. https://doi.org/10.1007/s10639-024-12677-0

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., Van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., Chastain, C., Zic, J., Horst, S., Turker, I., Agarwal, R., Osmundson, E., Idrees, K., Kieman, C., Padmanabhan, C., Bailey, C., Schlegel, C., Chambless, L., Gibson, M., Osterman, T., & Wheless, L. (2023). *Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model* (Publication no: PMC10002821). PubMed Central. https://doi.org/10.21203/rs.3.rs-2566942/v1

Karaduman, H., & Yetişensoy, O. (2023). *Sosyal bilgiler ve yapay zekâ: Eğitimde ideal bir sentez mi?* [Social studies and artificial intelligence: An ideal synthesis in education?]. In A. F. Ersoy & H. Karaduman (Eds.), *Sosyal bilgilerde güncel okumalar 3* [Contemporary readings in social studies 3] (pp. 249–277). Eğiten Kitap.

Kee, T., Kuys, B., & King, R. (2024). Generative artificial intelligence to enhance architecture education to develop digital literacy and holistic competency. *Jarina*, *3*(1), 24-41. https://doi.org/10.24002/jarina.v3i1.8347

Kim, J., & Lee, Y. (2024). Accuracy evaluation of tree images created using generative artificial intelligence. *Journal of Digital Landscape Architecture, 9,* 1029-1037. https://doi.org/10.14627/537752098

Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*, *30*, 1265-1300. https://doi.org/10.1007/s10639-024-12878-7

Kıyak, Y.S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal*, *2024,* qgae065. https://doi.org/10.1093/postmj/qgae065

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J.M. (2024). AI literacy and its implications for prompt engineering strategies. *Computer and Education: Artificial Intelligence*, *6*, 100225. https://doi.org/10.1016/j.caeai.2024.100225

Koga, S., & Du, W. (2024). ChatGPT's limited accuracy in generating anatomical images for medical education. *Skeletal Radiology, 53*, 1595-1596. https://doi.org/10.1007/s00256-024-04655-x

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-63. https://doi.org/10.1016/j.jcm.2016.02.012

Kruse, C., Schneider, J., & Seeber, I. (2023). Validity claims in children-AI discourse: Experiment with ChatGPT. In O. Poquet, A. Ortega-Arranz, O. Viberg, I. Chounta, B. McLaren, & J. Jovanovic (Eds.), *Proceedings of the 16th International Conference on Computer Supported Education - Volume 1* (pp. 289-296). SciTePress. https://doi.org/10.5220/0012552300003693

Kuşçu, O., Pamuk, A.E., Sütay Süslü N., & Hosal, S. (2023). Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Frontiers in Oncology, 13*, 1256459. https://doi.org/10.3389/fonc.2023.1256459

Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*, *17*(5), 926-931. https://doi.org/10.1002/ase.2270

Lee, U., Han, A., Lee, J., Lee, E., Kim, J., Kim, H., & Lim, C. (2023). Prompt Aloud!: Incorporating image-GenAI into STEAM class with learning analytics using prompt data. *Education and Information Technologies*, *29*, 9575-9605. https://doi.org/10.1007/s10639-023-12150-4

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL·E 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, *8*, 100648. https://doi.org/10.1016/j.ssaho.2023.100648

Liao, J., Chen, X., Fu, Q., Du, L., He, X., Wang, X., Han, S., & Zhang, D. (2024). Text-to-image generation for abstract concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(4), 3360-3368. https://doi.org/10.1609/aaai.v38i4.28122

Lim, B., Seth, I., Kah, S., Sofiadellis, F., Ross, R.J., Rozen, W.M., & Cuomo, R. (2023). Using generative artificial intelligence tools in cosmetic surgery: A study on rhinoplasty, facelifts, and blepharoplasty procedures. *Journal of Clinical Medicine*, *12*, 6524. https://doi.org/10.3390/jcm12206524

Lim, Y., Choi, H., & Shim, H. (2025). evaluating image hallucination in text-to-image generation with question-answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(25), 26290-26298. https://doi.org/10.1609/aaai.v39i25.34827

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Mack, K.A., Qadri, R., Denton, R., Kane, S.K., & Bennett, C.L. (2024). "They only care to show us the wheelchair": Disability representation in text-to-image AI models. In F. Mueller, P. Kyburz, J. Williamson, C. Sas, M. Wilson, P. Dugas, & I. Shklovski (Eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1-23). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642166

Mackey, B.P., Garabet, R., Maule, L., Tadesse, A., Cross, J., & Weingarten, M. (2024). Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. *Discovering Artificial Intelligence, 4*(38), 1-5. https://doi.org/10.1007/s44163-024-00135-2

Magomere, J., Ishida, S., Afonja, T., Salama, A., Kochin, D., Yuehgoh, F., Hamzaoui, I., Sefala, R., Alaagib, A., Dalal, S., Marchegiani, B., Semenova, E., Crais, L., & Hall, S. M. (2025). The world wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. In J. W. Vaughan, S. Mohamed, S. Fazelpour, & T. B. Gillis (Eds.), *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)* (pp. 246-282). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732019

Ministry of National Education [MoNE]. (2018). *Sosyal bilgiler öğretim programı: İlkokul ve ortaokul 4,5,6, ve 7. sınıflar* [Social studies curriculum: Elementary and middle school 4th, 5th, 6th and 7th grades]. Author.

Montenegro Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J.M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, *12*, 153. https://doi.org/10.3390/computers12080153

Morton, J.L. (2024). On inscription and bias: data, actor network theory, and the social problems of text-to-image AI models. *AI Ethics*, 5, 775-790. https://doi.org/10.1007/s43681-024-00431-8

Nasution, N.E.A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, 2(1), em002. https://doi.org/10.29333/agrenvedu/13071

Nielsen, J.P.S., Buchwald, C.V., & Grønhøj, C. (2023). Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. *Acta OtoLaryngologica*, *143*(9), 779-782, https://doi.org/10.1080/00016489.2023.2254809

Nguyen, H., Nguyen, V., López-Fierro, S., Ludovise, S., & Santagata, R. (2024). Simulating climate change discussion with large language models: Considerations for science communication at scale. In D. Joyner (Ed.), *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)* (pp. 28–38). Association for Computing Machinery. https://doi.org/10.1145/3657604.3662033

Ogunleye, B., Zakariyyah, K.I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, *14*, 636. https://doi.org/10.3390/educsci14060636

OpenAI. (2023). *DALL·E 3* [Large language model]. https://openai.com/index/dall-e-3/

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Peled, T., Sela, H.Y., Weiss, A., Grisaru-Granovsky, S., Agrawal, S., & Rottenstreich, M. (2024). Evaluating the validity of ChatGPT responses on common obstetric issues: Potential clinical applications and implications. *International Journal of Gynecology & Obstetrics, 166*(3), 1127-1133. https://doi.org/10.1002/ijgo.15501

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*, 887. https://doi.org/10.3390/healthcare11060887

Sanusi, I. T., Oyelere, S. S., & Omidiora, J. O. (2022). Exploring teachers' preconceptions of teaching machine learning in high school: A preliminary insight from Africa. *Computers and Education Open, 3*, 100072. https://doi.org/10.1016/j.caeo.2021.100072

Seth, I., Lim, B., Cevik, J., Sofiadellis, F., Ross, R.J., Cuomo, R., & Rozen, W.M. (2024). Utilizing GPT-4 and generative artificial intelligence platforms for surgical education: An experimental study on skin ulcers. *European Journal of Plastic Surgery, 47*, 19. https://doi.org/10.1007/s00238-024-02162-9

Suppadungsuk, S., Thongprayoon, C., Krisanapan, P., Tangpanithandee, S., Garcia Valencia, O., Miao, J., Mekraksakit, P., Kashani, K., & Cheungpasitporn, W. (2023). Examining the validity of Chatgpt in identifying relevant nephrology literature: Findings and implications. *Journal of Clinical Medicine*, *12*, 5550. https://doi.org/10.3390/jcm12175550

Temsah, M.H., Alhuzaimi, A.N., Almansour, M. Aljamaan, F., Alhasan, K., Batarfi, M.A., Altamimi, I., Alharbi, A., Alsuhaibani, A.A., Alwakeel, L., Alzahrani, A.A., Alsulaim, K.B., Jamal, A., Khayat, A., Alghamdi, M.H., Halwani, R., Khan, M.K., Al-Eyadhy, A., & Nazer, R. (2024). Art or Artifact: Evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. *Journal of Medical Systems, 48*, 54. https://doi.org/10.1007/s10916-024-02072-0

Vartiainen, H., & Tedre, M. (2023). Using artificial intelligence in craft education: Crafting with text to-image generative models. *Digital Creativity*, *34*, 1-21. https://doi.org/10.1080/14626268.2023.2174557

Vartiainen, H., Kahila, J., Tedre, M., López-Pernas, S., & Pope, N. (2024). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241252820

Vartiainen, H., Tedre, M., & Jormanainen, I. (2023). Co-creating digital art with GenAI in K-9 education: Socio- material insights. *International Journal of Education Through Art*, *19*(3), 405-423. https://doi.org/10.1386/eta_00143_1

Wang, J., Liu, X. G., Di, Z., Liu, Y., & Wang, X. E. (2023). *T2IAT: Measuring valence and stereotypical biases in text-to-image generation* (Publication no: 2306.00905). Arxiv. https://arxiv.org/abs/2306.00905

Wang, N., Wang, X., & Su, Y.S. (2024). Critical analysis of the technological affordances, challenges and future directions of Generative AI in education: A systematic review, *Asia Pacific Journal of Education*, *44*(1), 139-155. https://doi.org/10.1080/02188791.2024.2305156

Wei, Q., Yao, Z., Cui, Y., Wei, B., Jin, Z., & Xu, X. (2024). Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *Journal of Biomedical Informatics*, *151*, 104620. https://doi.org/10.1016/j.jbi.2024.104620

Wittmann, J. (2023). Science fact vs science fiction: A ChatGPT immunological review experiment gone awry. *Immunology Letters*, *256-257*, 42-47. https://doi.org/10.1016/j.imlet.2023.04.002

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2024). *ImageReward: Learning and evaluating human preferences for text-to-image generation* (Publication no: 2304.05977). Arxiv. https://arxiv.org/pdf/2304.05977

Yang, S., & Chang, M.C. (2024). The assessment of the validity, safety, and utility of ChatGPT for patients with herniated lumbar disc A preliminary study. *Medicine Open*, *103*(23), 1-4. https://doi.org/10.1097/MD.0000000000038445

Yeh, H.C. (2024). Revolutionizing language learning: Integrating generative AI for enhanced language proficiency. *Educational Technology & Society*, *27*(3), 335-353. https://doi.org/%2010.30191/ETS.202407_27(3).TP01

Yeşilyurt, S., Dündar, R., Demir, R. Z. & Yeşilyurt, A. G. (2025). Teaching social studies with generative artificial ıntelligence tools: advantages and disadvantages. *Journal of Interdisciplinary Educational Research, 9*(20), 1-16. https://doi.org/10.57135/jier.1594253

Yetişensoy, O. (2025). Beyond algorithms: Unveiling the intersection of AI ethics, citizenship education, and social studies. In J. Zajda & A. Rapoport (Eds.), *Discourses of globalisation and citizenship education. globalisation, comparative education and policy research* (pp. 151-166). Springer. https://doi.org/10.1007/978-3-031-86145-1_9

Yetişensoy, O., & Karaduman, H. (2024). The effect of AI-powered chatbots in social studies education. *Education and Information Technologies*, *29*, 17035-17069. https://doi.org/10.1007/s10639-024-12485-6

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*, *10*(2), e24289. https://doi.org/10.1016/j.heliyon.2024.e24289

Zack T, Lehman E, Suzgun M., Rodriguez, J.A., & Celi, L.A. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, *6*, e12-22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zalzal, H.G., Cheng, J., & Shah, R.K. (2023). Evaluating the current ability of ChatGPT to assist in professional potolaryngology education. *Oto Open*, *7*(4), e94. https://doi.org/10.1002/oto2.94

Zarei M., Zarei, M., Hamzehzadeh, S., Oliyaei S.S.B., & Hosseini, M. (2024). ChatGPT, a friend or a foe in medical education: A review of strengths, challenges, and opportunities. *Shiraz E-Medical Journal, 25*(6), e145840. https://doi.org/10.5812/semj-145840

Zhang, E.Y., Cheok, A.D., Pan, Z., Cai, J., & Yan, Y. (2023). From Turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. *Sci*, *5*, 46. https://doi.org/10.3390/sci5040046

Zhang, L., Liao, X., Yang, Z., Gao, B., Wang, C., Yang, Q., & Li, D. (2024). Partiality and misconception: Investigating cultural representativeness in text-to-image models. In F. Mueller, P. Kyburz, J. Williamson, C. Sas, M. Wilson, P. Dugas, & I. Shklovski (Eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1-25) Association for Computing Machinery. https://doi.org/10.1145/3613904.3642877

Appendix 1. Main category/theme, generic categories, and codes in the context of qualitative analysis