Art or Artifact: Evaluating the Accuracy, Appeal, and Educational Value of AI-Generated Imagery in DALL·E 3 for Illustrating Congenital Heart Diseases

Mohamad-Hani Temsah (mtemsah@ksu.edu.sa)
King Saud University, Riyadh
Abdullah N Alhuzaimi
King Saud University, Riyadh
Mohammed Almansour
King Saud University, Riyadh
Fadi Aljamaan
King Saud University, Riyadh
Khalid Alhasan
King Saud University, Riyadh
Munirah A Batarfi
King Abdullah International Medical Research Center, Riyadh
Ibraheem Altamimi
King Saud University, Riyadh
Amani Alharbi
King Saud University Medical City, King Saud University, Riyadh
Adel Abdulaziz Alsuhaibani
King Saud University Medical City, King Saud University, Riyadh
Leena Alwakeel
King Saud University Medical City, King Saud University, Riyadh
Abdulrahman Abdulkhaliq Alzahrani
King Saud University, Riyadh
Khaled B. Alsulaim
King Saud University, Riyadh
Amr Jamal
King Saud University, Riyadh
Afnan Khayat
Prince Sultan Military College of Health Sciences, Al Dharan
Mohammed Hussien Alghamdi
King Saud University, Riyadh
Rabih Halwani

University of Sharjah, Sharjah
Muhammad Khurram Khan
King Saud University, Riyadh
Ayman Al-Eyadhy
King Saud University, Riyadh
Rakan Nazer
King Saud University, Riyadh

Research Article

Keywords: AI-Generated Imagery, AI Text-to-Image generator, Congenital Heart Diseases, DALL·E 3 and Medical Education, Anatomical Accuracy, Healthcare Professional Visual Perceptions, Medical Illustrations and ChatGPT Integration, Medical artificial intelligence

Posted Date: January 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3895175/v1

Additional Declarations: No competing interests reported.

Abstract

Artificial Intelligence (AI), particularly AI-Generated Imagery, holds the capability to transform medical and patient education. This research explores the use of AI-generated imagery, from text-to-images, in medical education, focusing on congenital heart diseases (CHD). Utilizing ChatGPT's DALL·E 3, the research aims to assess the accuracy and educational value of AI-created images for 20 common CHDs. The study involved generating a total of 110 images for normal human heart and 20 common CHDs through DALL·E 3. Then, 33 healthcare professionals systematically assessed these AI-generated images by variable levels of healthcare professionals (HCPs) using a developed framework to individually assess each image anatomical accuracy, in-picture text usefulness, image appeal to medical professionals and the potential to use the image in medical presentations. Each item was assessed on a Likert scale of three. The assessments produced a total of 3630 images' assessments. Most AI-generated cardiac images were rated poorly as follows: 80.8% of images were rated as anatomically incorrect or fabricated, 85.2% rated to have incorrect text labels, 78.1% rated as not usable for medical education. The nurses and medical interns were found to have a more positive perception about the AI-generated cardiac images compared to the faculty members, pediatricians, and cardiology experts. Complex congenital anomalies were found to be significantly more predicted to anatomical fabrication compared to simple cardiac anomalies. There were significant challenges identified in image generation. These findings suggest adopting a cautious approach in integrating AI imagery in medical education, emphasizing the need for rigorous validation and interdisciplinary collaboration. The study advocates for future AI-models to be fine-tuned with accurate medical data, enhancing their reliability and educational utility.

Introduction

Illustrations and images are powerful methods to convey rich information and are widely used in medical practice [1]. The saying "a picture is worth a thousand words" appropriately highlights the value of medical illustrations in effectively conveying information to healthcare professionals and patients. This principle emphasizes the role of visual aids in simplifying complex medical concepts, making them more understandable and impactful. In instructional design, it is established that images enhance learning, a concept supported by literature [2–4]. This enhancement is supported by the mental model theory, which advocates that text and pictures facilitate the creation of both verbal (propositional) and visual mental models [5–7]. These models are then integrated into the learner's working memory as an aid in understanding and smooth future retrieval.[6] Images are generally considered less cognitively demanding than text. Text needs to be interpreted into concepts and then into a mental model, whereas images directly assist in creating a mental model due to their visual nature [8].

AI-powered text-to-image generators (AI-TIG) hold promise for medical illustrations, optimizing the self-learning principles like self-determination theory, adult learning theory, and the experiential learning cycle [9, 10]. These tools cater to learners' motivation and autonomy, aligning with adult learning's self-directed nature and experiential learning's emphasis on a four-stage cycle, namely the concrete experience, reflective observation, abstract conceptualization, and active experimentation, which can be perfectly

applied to the AI-TIG medical images and scenarios for training [11, 12]. AI-TIG can also create realistic and interactive simulations of medical situations, such as surgeries, emergencies, or clinical scenarios, that can help students and practitioners to learn and practice their skills and knowledge [13].

OpenAI announced DALL·E, a deep learning model, on January 5, 2021 [14]. It is a transformer-based model trained to generate images from text prompts. In 2023, AI-TIG applications, like DALL·E 3 and Midjourney, had significant advancement, creating better-detailed images [15, 16]. DALL·E 3, more detailed than its predecessor DALL·E 2, translates words into vibrant images and integrates with ChatGPT-4 [16, 17].

In medicine, previous AI-TIG models, like DALL·E 2, have shown potential, such as in the field of radiology [17]. These tools generated “realistic” x-ray images from text prompts and were seen as promising for image augmentation and manipulation in healthcare. However, their capabilities in generating specific images, such as CT, MRI, or ultrasound, or the abilities for generating images with pathological abnormalities, like fractures or tumors, remained limited [17]. There is a growing interest in exploring how these tools can be fine-tuned and adapted for medical applications [18, 19].

While previous studies investigated using deep learning, specifically neural network, to model cardiac anatomies representing the various types of Congenital Heart Diseases (CHD) and heart shape variations in cardiac disease, however, none of previous studies had in-depth evaluation about the educational value of the widely-available deep learning AI-TAG of DALL·E 3 [20–26]. We aimed to investigate the effectiveness and perfection of DALL·E 3 in producing educational illustrations for medical education, with a focus on CHD.

The study evaluated the accuracy and educational value of AI-TIG images for 20 common CHDs. Additionally, we explored the medical professionals' and students' perceptions of the utility and visual appeal of these AI-generated images in an educational context.

Methodology

Study design:

Our model evaluation study investigated the tendency of DALL·E 3 to generate scientifically accurate versus fictional images of common heart lesions. We conducted the text-to-pictures generative experiments with prompts designed to resemble a hypothetical potential usage by medical students or general healthcare providers of DALL·E 3 within clinical and medical education applications, taking the examples of CHDs (Appendix-1).

Selection of CHDs:

In the first phase of our study, we identified the most relevant CHDs for educational purposes. This was achieved through the expertise of two proficient pediatric cardiology experts (Drs. AAH and MAG). They

compiled a comprehensive list of top 20 CHDs that they frequently discussed in their educational sessions. This list (Appendix-1) served as the foundation for the subsequent AI-TIG process.

Prompt Optimization and Selection Strategy:

This phase focused on choosing the most effective prompts for generating illustrative images of CHDs, to ensure the reproducibility and educational relevance of the AI-generated images. It involved:

1. Pilot Testing of Various Prompts: We experimented with different prompt structures to ensure these prompts produced similar images. Examples of the prompts we tried:
    a. “Draw a 2D accurate illustration of [CHD] to simplify it for medical students, with text in the image to clarify the illustration.”
    b. “Draw an accurate 3D illustration of [CHD] to simplify it for medical students, with text in the image to clarify the illustration.”
    c. “Draw an accurate illustration of [CHD] like those in the Congenital Heart Disease: A Diagrammatic Atlas by Mullins and Mayer.” [27]
    d. “Draw a black and white accurate illustration of [CHD] to simplify it for medical students, with text in the image to clarify the illustration.”
2. A unique 'reverse engineering' approach was also employed. Here, we uploaded actual CHD illustrations into DALL·E 3, allowing the AI-TIG to describe them. The same text was then used to generate new images of the same CHD. This method helped in enhancing the prompt strategy by optimizing its text to match DALL·E 3 expectations and algorithm as much as possible.
3. Expert Panel Evaluation: A panel of medical experts reviewed the images from these various prompts.
4. Consistency Analysis: We assessed visual similarities of images produced from different prompts.
5. Final Prompt Selection: The chosen prompt template (as described below) was chosen by the expert panel as those that would be more likely used by medical students, healthcare providers or laypersons seeking illustration of CHD in AI-TIG (DALL·E 3).

Generation of Illustrative Images:

The creation of illustrative images was conducted using ChatGPT-4 integrated with DALL·E 3, under the supervision of the principal investigator, Dr. MHT. Over the course of three consecutive days, from November 29 to December 1, 2023, a series of prompts were issued to generate “accurate and educationally useful” illustrations based on the above-described methodology. MHT used the prompts in ChatGPT-4 as follows: “Draw an accurate illustration of]CHD[to simplify it for medical students, with text in the image to clarify the illustration” (Appendix-1). The aim was to produce a range of visual representations for each CHD, with five repetitions for each. Ten images of a normal heart were also generated to establish baseline for comparison, with the following prompt “Draw an accurate illustration

of a normal human heart to simplify it for medical students, with text supported image to clarify the illustration."

Development of the Image Assessment Framework:

A key component of our study was the development of a robust systemic framework for assessing the generated images. To accomplish this, an interdisciplinary expert panel was assembled, including two pediatric cardiologists (AAH, MAG), a cardiac surgeon (RN), an anatomist (MB), a medical educator (MAM), and two pediatricians (MHT, AAE). The panel developed a concise yet comprehensive evaluation tool, focusing on four key parameters: anatomical accuracy, value of integrated image-text, visual appeal to medical professionals, and usefulness for educational usage. Each image was assessed against the following criteria:

- Image Accuracy (accurate (score 3), midway (score 2), fabricated (score 1)) compared to a predefined criteria of each CHD and a "gold standard image", described below.
- Image-text usefulness (useful (score 3), midway (score 2), useless (score 1))
- Attractiveness to medical professionals (attractive (score 3), midway (score 2), not attractive (score 1))
- Suitability for medical education (as is (score 3), after modification (score 2), not useful (score 1))
- Overall image perfection score: calculated by summing the 4 above criteria items, which ranges between 4 and 12.

Validation of the Assessment Tool:

Prior to its application, the assessment tool described above underwent a thorough review process involving all co-authors of the study. This was essential to ensure the clarity and face validity of the tool to all team members.

Images Review and Assessments:

For the review and assessment phase, an online interface was set up on SurveyMonkey (Appendix 1). This platform hosted the collection of 110 colored images (10 normal heart and 100 CHDs). The assessment criteria (Appendix 1) were also embedded in the data collection tool [28, 29]. Alongside each image, the assessment scale was provided. The assessors were granted one-time access to this data-assessment portal, where they employed the agreed-upon assessment tool to evaluate each image. This method facilitated efficient and systematic data collection.

Ethical Considerations:

The Institutional Review Board (IRB) granted the approval of the proposal (Ref. No. 23/0155/IRB), and informed consent was obtained from the evaluators before their voluntary participation.

Statistical Analysis:

The mean and standard deviation were used to describe continuous variables and the frequencies and/ percentages for the categorically measured variables. The ratings of images were transformed from long data into wide data to account image sequence in the analysis, the resulted data matrix was equal to (110 image ratings*33 raters = 3630 image rating lines). The Cronbach's alpha test was applied to assess the internal consistency of the four measured cardiac image ratings or perceptions. The chi-squared test of association was used to assess the associations between categorically measured variables and the Spearman's (rho) correlations test was used to assess correlations between ordinal measured variables. The Spearman's Rho correlations test was used to assess the correlations between metric variables. A total relevance score for the AI generated images was computed via summing up the four indicators that characterized the images quality. These include following the four domains: anatomical accuracy, text usefulness, attractivity and usability for medical purposes.

The Generalized Liner Mixed Modelling with Gamma regression and Loglink was applied to evaluators mean overall AI-generated cardiac anomalies images perfection via regressing it against rater's demographic and professional characteristics with CHD complexity classifications. The association between the predictor variables with the dependent outcome variable in the GLMixed modelling was expressed as a multivariate adjusted Risk Rate (exponentiated beta coefficient) with its associated 95% confidence intervals. The SPSS IBM statistical software version #28 was used for the statistical data analysis and alpha significance level was considered at 0.050 level.

Results

In the study, 33 HCPs evaluated 110 cardiac images produced by DALL·E 3. The group consisted of diverse medical experts: eight (24.2%) cardiology experts, including a cardiac surgeon, three pediatric cardiology consultants, three fellows, and an anatomy consultant. Others included seven pediatricians, four non-pediatric faculty members, ten trainees (three medical students, four interns, three pediatric residents), and four pediatric nurses. Using an online data collection tool, this varied cohort completed 3630 individual image assessments, providing a comprehensive analysis of the AI-generated imagery. The evaluators also rated each cardiac anomaly; whether it was considered as simple or complex (Figure-1).

Evaluators' Overall Rating of AI-TIG CHD Images:

The evaluators' overall ratings for the AI-TIG cardiac images (N = 3630 ratings) are shown in Figure-2. Very few of the images (2.5%) were considered anatomically accurate, 16.7% as midway, and the majority (80.8%) were assessed as fabricated. In the evaluation of images' text label, 85.2% were rated as useless, only 1.2% were considered useful, and 13.6% fell into a mid-range of usefulness.

Regarding images' attractiveness, evaluators rated 18.7% of images as attractive, 18.2% as midway attractive, but most of images (63.1%) were considered as "not attractive at all". When considering usefulness for medical education, 78.1% were rated as "non usable", 21.6% as usable after modifications, while only 0.4% were evaluated as usable without modification.

Variation of rating of AI-TIG Cardiac Images among various evaluator groups:

The rating of images regarding the four different domains (anatomical accuracy, text usefulness, attractivity, usefulness for medical education) were compared among different groups of evaluators using the chi-squared test (Table 1). The medical students/interns/residents were found to be significantly more predicted to perceive the images as anatomically accurate, the illustrative text as useful, usable for medical educational purposes and attractive compared to the rest of evaluators (p-value < 0.001).

Likewise, nurses perceived the images significantly more compared to others as attractive, useful for medical education and its illustrative text as useful (p-value < 0.001). Conversely, the cardiology experts were significantly more inclined to perceive the images as (inaccurate, not attractive, not for medical education and their illustrative text being not useful) compared to the other evaluators.

Table 1
Evaluators' ratings of the AI-generated cardiac images (anatomical accuracy, text usefulness, attractiveness, usefulness for medical education). N = 3630 image ratings.

	Cardiac Images Anatomic Accuracy			
	Accurate %	Midway %	Fabricated %	p-value
Cardiology experts	12.3	10.1	31.2	< 0.001
Pediatrics specialist/ consultant	0	22.9	17.8	
Faculty member	12.2	12	12.1	
Medical interns/students/residents	75.5	33.5	28.3	
Nurses	0	21.5	10.6	
	Illustrative text usefulness for viewer			
	Usefulness for viewer %	Midway %	Not useful %	p-value
Cardiology experts	7.1	3.6	31.3	< 0.001
Pediatrics specialist/ consultant	0	18.3	18.5	
Faculty member	0	4.8	13.4	
Medical interns/students/residents	50	30.5	30	
Nurses	42.9	42.8	6.8	
	Useability for medical education			
	Useable as is %	after modification %	Not useable %	p-value
Cardiology experts	7.7	6	33.2	< 0.001
Pediatricians (specialist/consultants)	0	14.6	19.3	
Faculty member	0	10.9	12.5	
Medical interns/students/residents	76.9	39.3	27.6	
Nurses	15.4	29.2	7.4	
	Attractivity			

	Cardiac Images Anatomic Accuracy			
	Attractive %	Midway %	Not attractive %	p-value
Cardiology experts	4.1	10.4	39	< 0.001
Pediatricians (specialist/consultants)	8.4	19.3	20.8	
Faculty member	7.7	16.3	12.2	
Medical interns/students/residents	52.8	35	22.3	
Nurses	27	19	5.7	

Rating AI-TIG cardiac images of normal hearts, simple and complex CHD lesions:

The AI-TIG images of normal hearts (Figure-3) were rated poor regarding anatomic accuracy (47.9% fabricated, 40.3% midway and only 11.8% accurate). An example of the “most fabricated images” is shown in Figure-4a, and “least fabricated” in Figure-4b. Moreover, 83.9% of images of normal heart were rated as having inaccurate and useless text labels. In addition, 64.2% of images of them were rated as not useable for medical education, 34.5% can be used after modification, and only 1% thought these images can be used without modification.

This extends to the individual rating of the AI-TIG images of the various CHDs that have been studied. Most AI-TIG images were rated poor regarding anatomical accuracy, illustrative text usefulness and usability for medical education 1–3%. However, generally the images were perceived as attractive in 15–22%.

Chi-squared test (Table S1) showed that the CHD complexity correlated significantly with the evaluators’ perceived images’ anatomical accuracy. Complex CHD images were found to be significantly more fabricated compared to normal heart or simple CHD, p-value < 0.001. While the other three evaluation criteria (image’s text usefulness, attractiveness, or usefulness for medical education) did not significantly correlate with CHD complexity.

Correlations between evaluators’ perceptions of the four criteria of AI-TIG Cardiac Anomalies Images:

Table 2 highlights the bivariate correlation between the four-criterion used to assess images quality. We found significantly positive correlation (*P = 0.01*) between all of them (r ranged between 0.337–0.566). The best correlation was between image usefulness for medical education and its attractiveness. Furthermore, the lowest correlation was between image attractiveness and its anatomic accuracy.

Usefulness for medical education overall had the best correlation with all the other three criteria (r ranged between 0.441–0.566).

Table 2
Bivariate Spearman's Correlations between evaluator's perceptions of the AI generated cardiac anomalies images.

	Anatomic Accuracy	Text labels Usefulness	Image Attractiveness
Anatomic Accuracy			
Text labels Usefulness	.394**		
Image Attractiveness	.337**	.388**	
Useability For Medical Education	.497**	.441**	.566**
** Correlation is significant at the 0.01 level (2-tailed).			

Multivariable Analysis of evaluators perceived overall perfection score of AI-TIG Cardiac Images:

We ran multivariable generalized linear regression for the overall mean perfection score of the AI-TIG cardiac anomalies images in comparison to cardiology experts mean perfection score. Nurses had significantly the highest perfection score compared to cardiology experts (34.1% times higher $p < 0.001$), followed by medical students/interns/residents (26.6% times higher $p < 0.001$), then faculty staff/academician (15.5% higher $p < 0.001$). Pediatric consultant/specialist had higher perfection score by 14.5% times higher $p < 0.001$).

Taking cardiac anomaly complexity into consideration (Table 3), complex ones were evaluated significantly less perfect compared to simple ones in overall by all evaluators (6% times less $p < 0.001$). For example, certain anomalies, like the coarctation of Aorta, Interruption of aortic, Aorto-left ventricular tunnel, were perceived significantly less perfect by all evaluators (4.4%-11% less perfect) as compared to other CHD images.

Table 3
Multivariable Generalized Linear Regression (GLM) analysis of evaluators perceived Overall Relevance of AI generated cardiac images score .

Model Term	Multivariate adjusted Risk Rate (RR)	95% CI for RR		
		Lower	Upper	p-value
(Intercept)	5.844	5.722	5.969	< 0.001
Clinical role = Nurses	1.341	1.306	1.378	< 0.001
Clinical role = Medical student/intern/residents	1.266	1.244	1.289	< 0.001
Clinical role = Faculty staff/Academician	1.155	1.128	1.183	< 0.001
Clinical orle = Pediatric consultant/specialist	1.145	1.122	1.170	< 0.001
Cardiac anomaly image complexity level = Complex	0.940	0.926	0.955	< 0.001
Cardiac anomaly image = Hypoplastic left heart syndrome (HLHS)	1.058	1.024	1.093	0.001
Cardiac anomaly image = Ebstein syndrome	0.969	0.938	1.002	0.067
Cardiac anomaly image = Coarctation of Aorta	0.889	0.860	0.919	< 0.001
Cardiac anomaly image = Interruption of aortic arch (IAA)	0.912	0.882	0.943	< 0.001
Cardiac anomaly image = AP window	0.956	0.925	0.988	0.008
Cardiac anomaly image= Aorto-left ventricular tunnel	0.933	0.903	0.965	< 0.001
Dependent Variable: Healthcare professional's mean perceived overall relevance of images score (sum of the 4 rating aspects)				

Discussion

DALL·E 3 is an AI-TIG model that generates images from text descriptions through transformative language models like GPT-3 [16, 30]. It can produce a variety of images, from realistic to abstract art, and can creatively combine elements from different ideas to create novel visuals. Despite its potential in areas like education and art, DALL·E faced challenges, such as generating coherent images from complex texts, maintaining image quality, addressing biases from training data, and managing computational demands [31].

AI-TIG models have shown proficiency in generating images with correct style and content for some medical applications, such as histopathology and scientific illustrations [32]. Some potential benefits of these technologies include educational applications without copyright limitations, tailored educational experience, data anonymization, and discovery of new morphological associations. Conversely, they have potential limitations that lie in their current inability to accurately generate complex medical images.

While offering innovative visual learning AI-tools, AI-TIG's integration in education requires careful balance and validation for accuracy and reliability, similar to Large Language Models (LLMs) [18, 33, 34]. We demonstrated that AI-TIG, like LLMs, is liable to generate inaccuracies ('hallucinations' or 'confabulation'), posing risks in medical contexts. Consequently, one recommended approach for AI-TIG use is the 'sandwich technique': experts input text, AI-TIG generates the image, and then the expert evaluates and edits it for accuracy, ensuring safer application in the educational process [35].

Our study explored the current state of DALL·E 3 in the field of medical illustrations, particularly CHDs. We discovered that while this technology opens novel avenues for visualizations, it also poses significant challenges. Like the "hallucinations" in LLMs, the tendency of DALL·E 3 to introduce inaccuracies and 'artifacts' in images was significant, raising concerns about its current suitability for medical illustrations [36]. These insights emphasize the need for rigorous validation before employing AI-TIG imagery in complex areas like medical education, patient's education, or decision-making.

Our study found that the majority of 3630 evaluations rated DALL·E 3's AI-generated cardiac images as anatomically inaccurate and educationally limited. These shortcomings may stem from the model's training and its 'Zero-Shot' ability, which inconsistently adapts to untrained text prompts [14, 37]. However, other research on AI has shown promise in enhancing medical imaging quality and interpretability in cardiology [1]. Despite DALL·E 3's current limitations, ongoing research and developments may improve AI-TIG medical images' accuracy.

Another concern in our study was the erroneous AI-generated images text-labels, that were mostly misspelled or misplaced, rendering them "useless". For enhanced medical illustrations, future AI-TIG models should be developed to meticulously produce accurate medical images labeling [1]. Specialized or fine-tuned GPT models could be trained to more accurately recognize medical structures and enhance their labeling [38]. As these AI-TIGs undergo more medically-oriented training, their accuracy may improve, providing a better learning and personalized medical tool for healthcare professionals, patients, and educators [39].

Interestingly, 18% images in our sample were thought of as having "attractive appearance" for medical professionals, as was also noted by other studies describing DALL·E 3 images as more realistic [15]. Nurses and junior trainees in our group had more positive perception about AI-TIG cardiac images; perceiving more images as anatomically "accurate," finding the illustrative text as more useful and usable for medical educational purposes and seeing more attractive images than the other evaluators. While these could be a positive signal for future medical curriculum adaptation of more accurate AI models,

these findings may indicate a risk of persuading non-expert medical professionals or laypersons to be influenced by the vibrant artistic appearance of such images.

In our study, AI-TIG cardiac images, including those of normal hearts and simple lesions, were frequently rated poorly in terms of anatomical accuracy. This issue may be attributed to inherent challenges in DALL·E's capabilities, including difficulties in image coherence, quality, and biases in training datasets [31]. Moreover, while complex congenital anomalies were more prone to anatomical fabrication, the complexity of cardiac disease did not significantly impact the perceived educational value of these images. Notably, there was a positive correlation between the perceived anatomical accuracy and educational usefulness of the images, emphasizing the importance of accuracy for medical education purposes.

The expert panel also observed additional inaccuracies in the AI-generated images, such as the depiction of non-existent blood vessels in the heart images and a notable lack of cardiac valves. In addition, the AI-model apparently did not seem to identify the various structures of the heart (e.g. aorta, pulmonary valve, atrial or ventricular septum), therefore, it could not draw the abnormalities of these structures neither link these structures to correct text labels. This is like several errors that were reported in the illustrations of the heart by three AI-TIGs: Microsoft Bing/DALL·E, Stable Diffusion and Craiyon [40]. The investigator used the prompt to draw “detailed and accurate anatomy illustration of the human heart” on the three platforms on May 30, 2023, and found that they failed to show accurate coronary artery origins, the branching of the aorta and pulmonary trunk.

The inaccuracy issues may stem from DALL·E 3 possibly being trained on unrepresentative data, leading to a risk of overfitting to inaccurate disease images from automation bias.[41] Sharing such flawed images and illustrations to non-cardiac experts, like medical students, nurses, or laypersons, could unintentionally generate or intensify misinformation, a concern exacerbated by automation biases. This highlights the need for caution in using AI-tools for didactic purposes, particularly in sensitive fields like healthcare education [42–47].

To mitigate some risks of AI-TIG medical imagery, it is important to educate HCPs and patients on proper use of AI tools, such as appropriate prompts that are more specific and at higher levels of medical literacy to produce higher-quality images [48]. Also, careful interpretation of the medical images still requires experts’ oversight, to ensure images are not misinforming users [35, 42]. One capability of AI-models is their ability to acquire knowledge and improve performance through increased exposure to data, therefore, IT experts could enhance current and future AI-models’ training, emphasizing variety of accurate medical images datasets and improving algorithms to enhance generated image’s reliability and usefulness in medical education [49, 50].

Medical digital twins, serving as virtual representations of medical conditions, could improve merging the physical and virtual medical realms [51]. Recently, digital twin technology, especially in cardiac modeling, witnessed substantial progress [52]. However, challenges of the variability of human heart parameters and their implications on patient response to treatments persist, and personalized digital twins that

mimic specific heart pathologies demand significant computational resources [51]. Therefore, AI-TIG may offer new opportunities, provided these models are both accurate, widely accessible, and easily editable, thus improving the personalized healthcare provision and medical education experience, in various medical fields [53, 54].

Study Limitations and Future Potentials:

Our study focused on one category of anatomical lesions (CHDs) at specific time on one AI-TIG (DALL·E 3). Therefore, future research of AI-TIG images for other health-related conditions or other AI-TIG models may produce variable outcomes. Our research is among the first to explore AI-TIG images potentials of DALL·E 3 in CHD, and it may pave the way for more medical-specific AI-training for future models.

Future research on AI-TIG may address other shortcomings, such as the 'black-box' nature of the models, the requirement for extensive medical-data training effects, better transparency of image standardization, or improved filtering of inaccuracies during training [55]. The optimal use of AI-TIG images in medical education or individualized healthcare with digital twin models requires further collaboration between healthcare professionals and computer scientists. This includes defining clear objectives, choosing the optimal deep learning algorithms and datasets, and interpreting image results with a balanced, human-supervised perspective.

Conclusion

This study explored the integration of AI-TIG technology in medical illustrations, particularly for visualizing CHDs, highlighting a novel approach. Despite experts identifying errors and questioning the medical utility of AI-generated images, non-experts like medical students and nurses viewed them more positively. These results point out the need for caution among AI-TIG users and healthcare professionals, emphasizing vigilance in their application. Additionally, there is an opportunity for computer scientists and AI stakeholders to refine AI-TIG models with more realistic medical images. Importantly, text in the AI-TIG generated images should clearly indicate potential inaccuracies in both visuals and descriptions. Further research into other healthcare imaging techniques using generative AI is warranted.

Abbreviations

AI
Artificial Intelligence
AI-TIG
AI-powered text-to-image generator
ASD
Atrial Septal Defect
CCTGA
Congenitally Corrected Transposition of the Great Arteries

CHD
Congenital Heart Diseases
DALL·E 3
DALL·E version 3
DILV
Double Inlet Left Ventricle
DORV
Double Outlet Right Ventricle
GLMM
Generalized Linear Mixed Model
HCP
Healthcare Professional
HLHS
Hypoplastic Left Heart Syndrome
LVEF
Left Ventricular Ejection Fraction
MAPCA
Major Aortopulmonary Collateral Arteries
PDA
Patent Ductus Arteriosus
VSD
Ventricular Septal Defect

Declarations

Acknowledgement: Authors are grateful for all data assessment volunteers, namely Abdullah Taha, Abdulrahman Senjab, Ashwathy Nair, Dilsy Devassy, Jawad Alzamil, Maha Mubarak Binfadel, Mohammad Derbas, Muhammed Elbeddawy, Naila Muhammed, Omar Temsah, Salma Abdalla, Sayed Belal Sayed, Sukanya Sudevan, Yazan Chaiah Salah Omar, Ziyad Zaid Alkathiri. Also, we are grateful to hudhudata.com for the data analysis and statistical support. The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for supporting this research (IFKSURC-1-XXXX). During the preparation of this work the authors used ChatGPT, an AI Chatbot developed by OpenAI (San Francisco, California, U.S.), in order to improve the readability and language of this work, without replacing researchers' tasks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding: No funding was received to assist with the preparation of this manuscript.

Data Availability Statement: The data of this research is available from the corresponding author upon reasonable request.

Declaration of interests: The authors declare they have no financial interests.

References

1. Olender, M.L.; de la Torre Hernández, J.M.; Athanasiou, L.S.; Nezami, F.R.; Edelman, E.R. Artificial intelligence to generate medical images: augmenting the cardiologist's visual clinical workflow. Eur Heart J Digit Health 2021, *2*, 539–544, doi:10.1093/ehjdh/ztab052.
2. Levie, W.H.; Lentz, R. Effects of text illustrations: A review of research. Ectj 1982, *30*, 195–232.
3. Filippatou, D.; Pumfrey, P.D. Pictures, titles, reading accuracy and reading comprehension: a research review (1973-95). Educational Research 1996, *38*, 259–291.
4. Kools, M.; van de Wiel, M.W.; Ruiter, R.A.; Kok, G. Pictures and text in instructions for medical devices: effects on recall and actual performance. Patient Educ Couns 2006, *64*, 104–111, doi:10.1016/j.pec.2005.12.003.
5. Johnson-Laird, P.N. *Mental models: Towards a cognitive science of language, inference, and consciousness*; Harvard University Press: 1983.
6. Mayer, R.E. Multimedia learning: Are we asking the right questions? Educational psychologist 1997, *32*, 1–19.
7. Schnotz, W.; Bannert, M. Influence of the type of visualization on the construction of mental models during picture and text comprehension. Zeitschrift fur Experimentelle Psychologie: Organ der Deutschen Gesellschaft fur Psychologie 1999, *46*, 217–236.
8. Ganier, F. Processing text and pictures in procedural instructions. Information Design Journal 2000, *10*, 146–153.
9. Ker, J. Teaching made easy: A manual for health professionals. Bmj 2000, *320*, 1677.
10. Deci, E.L.; Ryan, R.M. *Intrinsic motivation and self-determination in human behavior*, Springer Science & Business Media: 2013.
11. Kolb, A.Y.; Kolb, D.A. Learning styles and learning spaces: Enhancing experiential learning in higher education. Academy of management learning & education 2005, *4*, 193–212.
12. Mukhalalati, B.A.; Taylor, A. Adult Learning Theories in Context: A Quick Guide for Healthcare Professional Educators. J Med Educ Curric Dev 2019, *6*, 2382120519840332, doi:10.1177/2382120519840332.
13. Reed, J.M. Using Generative AI to Produce Images for Nursing Education. Nurse Educ 2023, *48*, 246, doi:10.1097/nne.0000000000001453.
14. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. 2021; pp. 8821–8831.
15. Black, J. DALL-E 3 vs Midjourney: which AI photo tool is better? Available online: https://www.pickfu.com/blog/dall-e-vs-midjourney/ (accessed on 6 Jan 2024).
16. OpenAI. DALL·E 3. Available online: https://openai.com/dall-e-3 (accessed on 6 Jan 2024).

17. Adams, L.C.; Busch, F.; Truhn, D.; Makowski, M.R.; Aerts, H.; Bressem, K.K. What Does DALL-E 2 Know About Radiology? J Med Internet Res 2023, *25*, e43110, doi:10.2196/43110.
18. Jamal, A.; Solaiman, M.; Alhasan, K.; Temsah, M.H.; Sayed, G. Integrating ChatGPT in Medical Education: Adapting Curricula to Cultivate Competent Physicians for the AI Era. Cureus 2023, *15*, e43036, doi:10.7759/cureus.43036.
19. Hajar, R. Medical illustration: art in medical education. Heart Views 2011, *12*, 83–91, doi:10.4103/1995-705x.86023.
20. Kong, F.; Stocker, S.; Choi, P.S.; Ma, M.; Ennis, D.B.; Marsden, A. SDF4CHD: Generative Modeling of Cardiac Anatomies with Congenital Heart Defects. In *ArXiv*; United States, 2023.
21. Beetz, M.; Corral Acero, J.; Banerjee, A.; Eitel, I.; Zacur, E.; Lange, T.; Stiermaier, T.; Evertz, R.; Backhaus, S.J.; Thiele, H.; et al. Interpretable cardiac anatomy modeling using variational mesh autoencoders. Front Cardiovasc Med 2022, *9*, 983868, doi:10.3389/fcvm.2022.983868.
22. Qiao, M.; Wang, S.; Qiu, H.; De Marvao, A.; O'Regan, D.P.; Rueckert, D.; Bai, W. CHeart: A Conditional Spatio-Temporal Generative Model for Cardiac Anatomy. IEEE Trans Med Imaging 2023, *Pp*, doi:10.1109/tmi.2023.3331982.
23. Campello, V.M.; Xia, T.; Liu, X.; Sanchez, P.; Martín-Isla, C.; Petersen, S.E.; Seguí, S.; Tsaftaris, S.A.; Lekadir, K. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. Front Cardiovasc Med 2022, *9*, 983091, doi:10.3389/fcvm.2022.983091.
24. Vieira, M.S.; Hussain, T.; Figueroa, C.A. Patient-specific image-based computational modeling in congenital heart disease: a clinician perspective. Journal of Cardiology and Therapy 2015, *2*, 436–448.
25. Tikenoğulları, O.Z.; Peirlinck, M.; Chubb, H.; Dubin, A.M.; Kuhl, E.; Marsden, A.L. Effects of cardiac growth on electrical dyssynchrony in the single ventricle patient. Comput Methods Biomech Biomed Engin 2023, 1–17, doi:10.1080/10255842.2023.2222203.
26. Biffi, C.; Cerrolaza, J.J.; Tarroni, G.; Bai, W.; de Marvao, A.; Oktay, O.; Ledig, C.; Le Folgoc, L.; Kamnitsas, K.; Doumou, G.; et al. Explainable Anatomical Shape Analysis Through Deep Hierarchical Generative Models. IEEE Trans Med Imaging 2020, *39*, 2088–2099, doi:10.1109/tmi.2020.2964499.
27. Mullins, C.E.; Mayer, D.C. Congenital heart disease: a diagrammatic atlas. *(No Title)* 1988.
28. Park, I.S. *An Illustrated Guide to Congenital Heart Disease: From Diagnosis to Treatment–From Fetus to Adult*; Springer: 2019.
29. Ottaviani, G.; Buja, L.M. Congenital heart disease: pathology, natural history, and interventions. In *Cardiovascular pathology*; Elsevier: 2022; pp. 223–264.
30. Singh, S. 9 Capabilities Of DALL-E That One Must Know. Available online: https://www.labellerr.com/blog/dall-e-everything-you-need-to-know/ (accessed on 6 Jan 2024).
31. AppMaster. Challenges and Limitations: Understanding DALL-E's Capabilities. Available online: https://appmaster.io/blog/challenges-and-limitations-dall-e (accessed on 6 Jan 2024).

32. Kather, J.N.; Ghaffari Laleh, N.; Foersch, S.; Truhn, D. Medical domain knowledge in domain-agnostic generative AI. NPJ Digit Med 2022, *5*, 90, doi:10.1038/s41746-022-00634-5.
33. Seetharaman, R. Revolutionizing Medical Education: Can ChatGPT Boost Subjective Learning and Expression? In *J Med Syst*; © 2023. The Author(s), under exclusive licence to Springer Science + Business Media, LLC, part of Springer Nature.: United States, 2023; Volume 47, p. 61.
34. BaHammam, A.S. Balancing Innovation and Integrity: The Role of AI in Research and Scientific Writing. In *Nat Sci Sleep*; New Zealand, 2023; Volume 15, pp. 1153–1156.
35. Temsah, R.; Altamimi, I.; Alhasan, K.; Temsah, M.H.; Jamal, A. Healthcare's New Horizon With ChatGPT's Voice and Vision Capabilities: A Leap Beyond Text. Cureus 2023, *15*, e47469, doi:10.7759/cureus.47469.
36. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus 2023, *15*, e35179, doi:10.7759/cureus.35179.
37. Simonsanvil. DALL-E-Explained. Available online: https://github.com/simonsanvil/DALL-E-Explained/blob/main/README.md (accessed on 6 Jan 2024).
38. OpenAI. Fine-tuning: Learn how to customize a model for your application. Available online: https://platform.openai.com/docs/guides/fine-tuning (accessed on 6 Jan 2024).
39. Temsah, M.H.; Jamal, A.; Aljamaan, F.; Al-Tawfiq, J.A.; Al-Eyadhy, A. ChatGPT-4 and the Global Burden of Disease Study: Advancing Personalized Healthcare Through Artificial Intelligence in Clinical and Translational Medicine. Cureus 2023, *15*, e39384, doi:10.7759/cureus.39384.
40. Noel, G. Evaluating AI-powered text-to-image generators for anatomical illustration: A comparative study. Anat Sci Educ 2023, doi:10.1002/ase.2336.
41. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012, *19*, 121–127, doi:10.1136/amiajnl-2011-000089.
42. Preiksaitis, C.; Rose, C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. JMIR Med Educ 2023, *9*, e48785, doi:10.2196/48785.
43. Liu, J.; Liu, F.; Fang, J.; Liu, S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023, *71*, 102064, doi:10.1016/j.outlook.2023.102064.
44. Kim, T.W. Application of artificial intelligence chatbot, including ChatGPT in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof 2023, *20*, 38, doi:10.3352/jeehp.2023.20.38.
45. Abdel Aziz, M.H.; Rowe, C.; Southwood, R.; Nogid, A.; Berman, S.; Gustafson, K. A scoping review of artificial intelligence within pharmacy education. Am J Pharm Educ 2023, 100615, doi:10.1016/j.ajpe.2023.100615.
46. Tiwari, A.; Kumar, A.; Jain, S.; Dhull, K.S.; Sajjanar, A.; Puthenkandathil, R.; Paiwal, K.; Singh, R. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. Cureus 2023, *15*, e40367, doi:10.7759/cureus.40367.

47. Padovan, M.; Cosci, B.; Petillo, A.; Nerli, G.; Porciatti, F.; Scarinci, S.; Carlucci, F.; Dell'Amico, L.; Meliani, N.; Necciari, G.; et al. ChatGPT in Occupational Medicine: A Comparative Study with Human Experts. Bioengineering 2024, *11*, doi:10.3390/bioengineering11010057.
48. Lautrup, A.D.; Hyrup, T.; Schneider-Kamp, A.; Dahl, M.; Lindholt, J.S.; Schneider-Kamp, P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. Open Heart 2023, *10*, doi:10.1136/openhrt-2023-002455.
49. Jone, P.-N.; Gearhart, A.; Lei, H.; Xing, F.; Nahar, J.; Lopez-Jimenez, F.; Diller, G.-P.; Marelli, A.; Wilson, L.; Saidi, A.; et al. Artificial Intelligence in Congenital Heart Disease: Current State and Prospects. JACC: Advances 2022, *1*, 100153, doi:https://doi.org/10.1016/j.jacadv.2022.100153.
50. Mohsin, S.N.; Gapizov, A.; Ekhator, C.; Ain, N.U.; Ahmad, S.; Khan, M.; Barker, C.; Hussain, M.; Malineni, J.; Ramadhan, A.; et al. The Role of Artificial Intelligence in Prediction, Risk Stratification, and Personalized Treatment Planning for Congenital Heart Diseases. Cureus 2023, *15*, e44374, doi:10.7759/cureus.44374.
51. Moztarzadeh, O.; Jamshidi, M.B.; Sargolzaei, S.; Jamshidi, A.; Baghalipour, N.; Malekzadeh Moghani, M.; Hauer, L. Metaverse and Healthcare: Machine Learning-Enabled Digital Twins of Cancer. Bioengineering (Basel) 2023, *10*, doi:10.3390/bioengineering10040455.
52. Viola, F.; Del Corso, G.; De Paulis, R.; Verzicco, R. GPU accelerated digital twins of the human heart open new routes for cardiovascular research. Sci Rep 2023, *13*, 8230, doi:10.1038/s41598-023-34098-8.
53. Corral-Acero, J.; Margara, F.; Marciniak, M.; Rodero, C.; Loncaric, F.; Feng, Y.; Gilbert, A.; Fernandes, J.F.; Bukhari, H.A.; Wajdan, A.; et al. The 'Digital Twin' to enable the vision of precision cardiology. Eur Heart J 2020, *41*, 4556–4564, doi:10.1093/eurheartj/ehaa159.
54. Peshkova, M.; Yumasheva, V.; Rudenko, E.; Kretova, N.; Timashev, P.; Demura, T. Digital twin concept: Healthcare, education, research. J Pathol Inform 2023, *14*, 100313, doi:10.1016/j.jpi.2023.100313.
55. Krittanawong, C.; Johnson, K.W.; Rosenson, R.S.; Wang, Z.; Aydar, M.; Baber, U.; Min, J.K.; Tang, W.H.W.; Halperin, J.L.; Narayan, S.M. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J 2019, *40*, 2058–2073, doi:10.1093/eurheartj/ehz056.

Figures

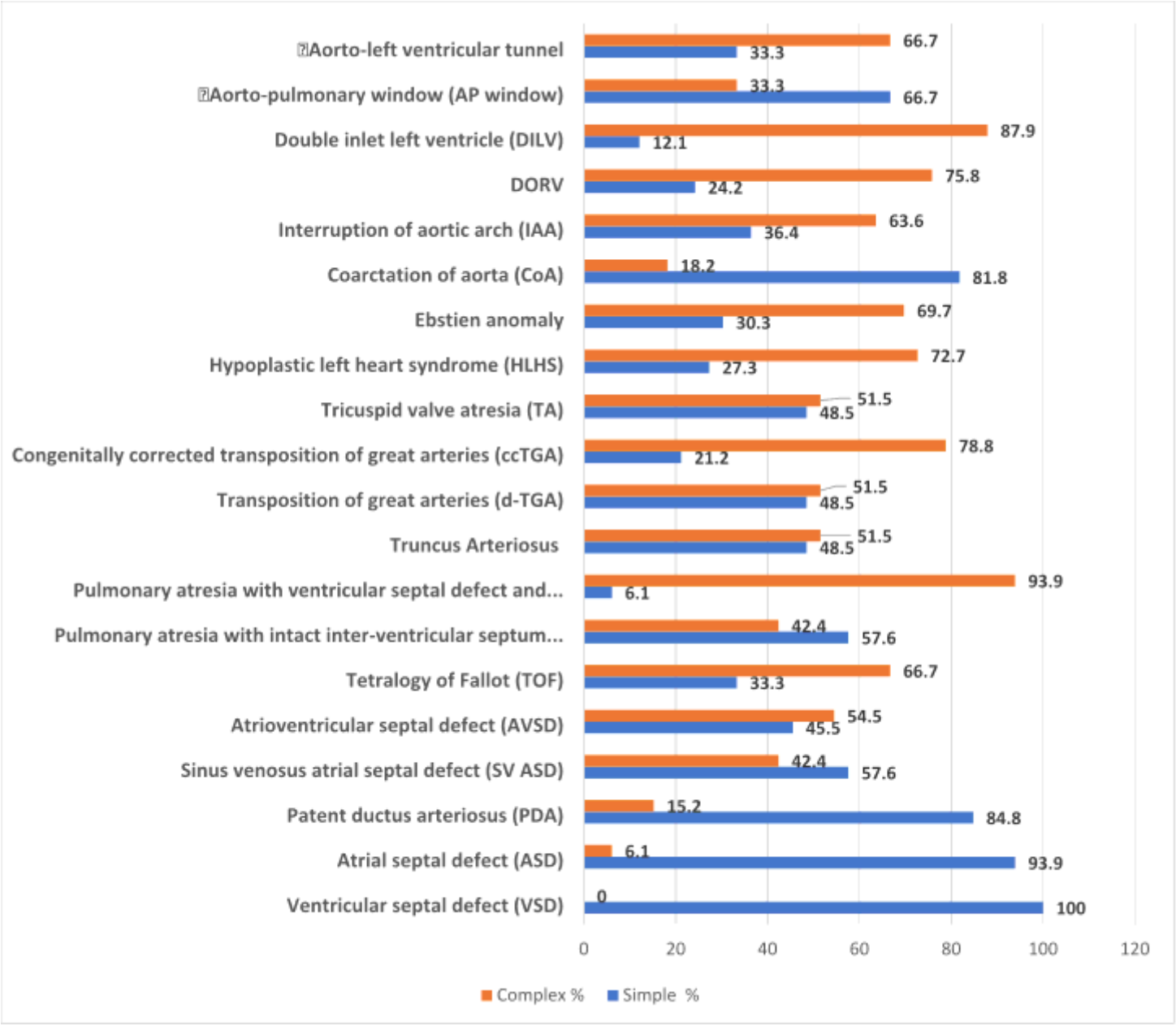

Figure 1

The evaluators' prior perceptions of the complexity of studied CHD anomalies, (CHD: Congenital Heart Disease)

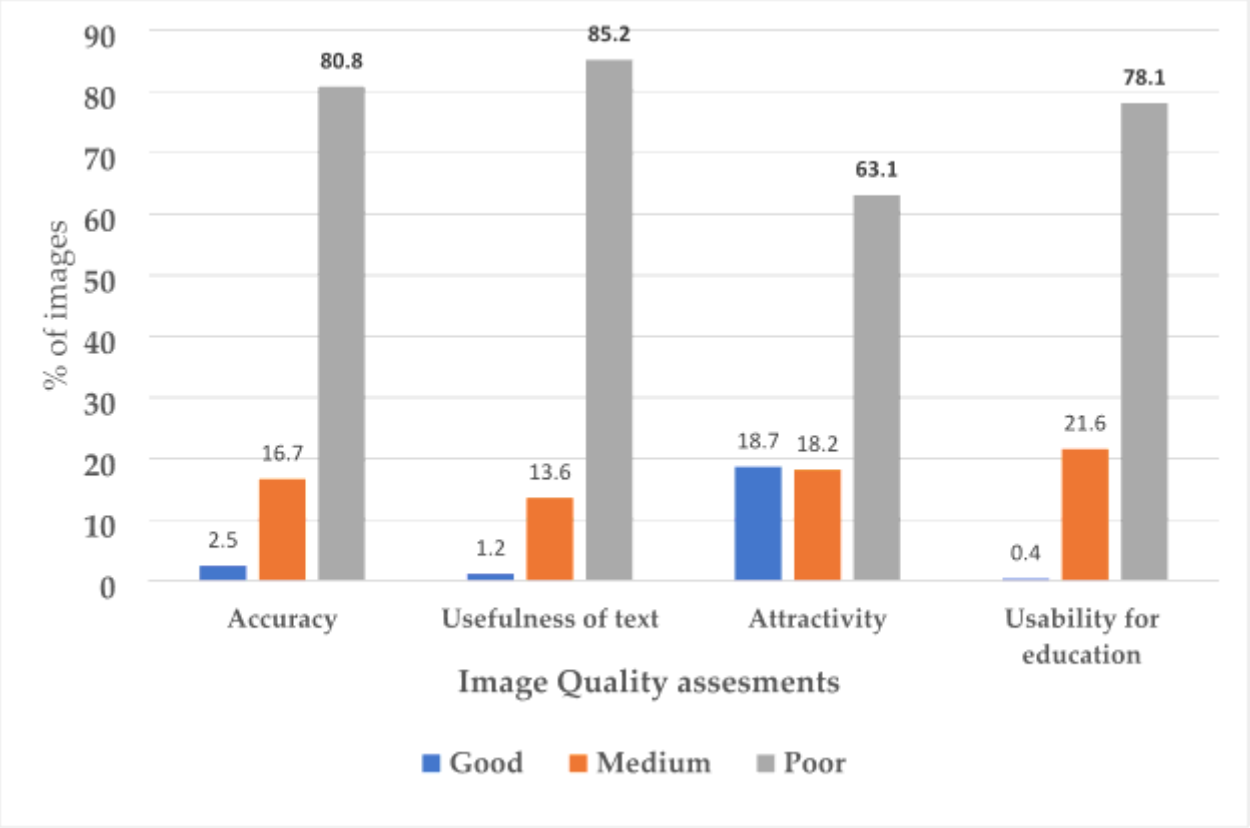

Figure 2

The Evaluators' overall rating of the AI-generated congenital cardiac anomalies images (N=3630 ratings)

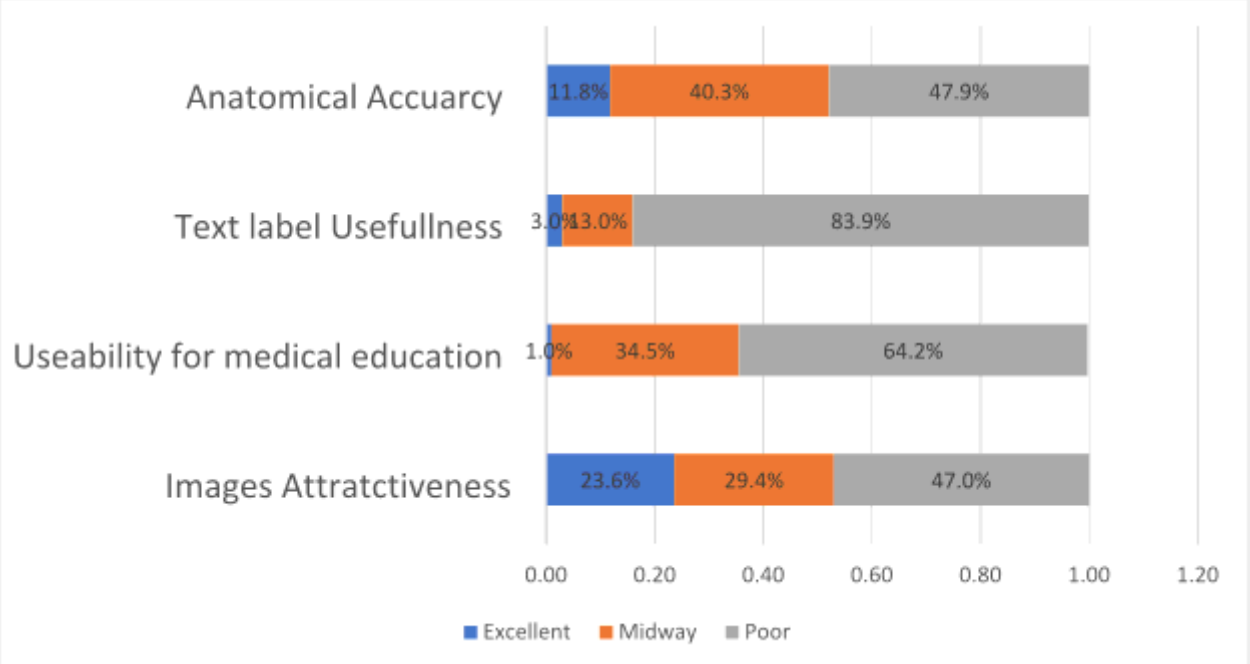

Figure 3

Evaluators' ratings for the AI-TIG Cardiac Images for Normal Heart

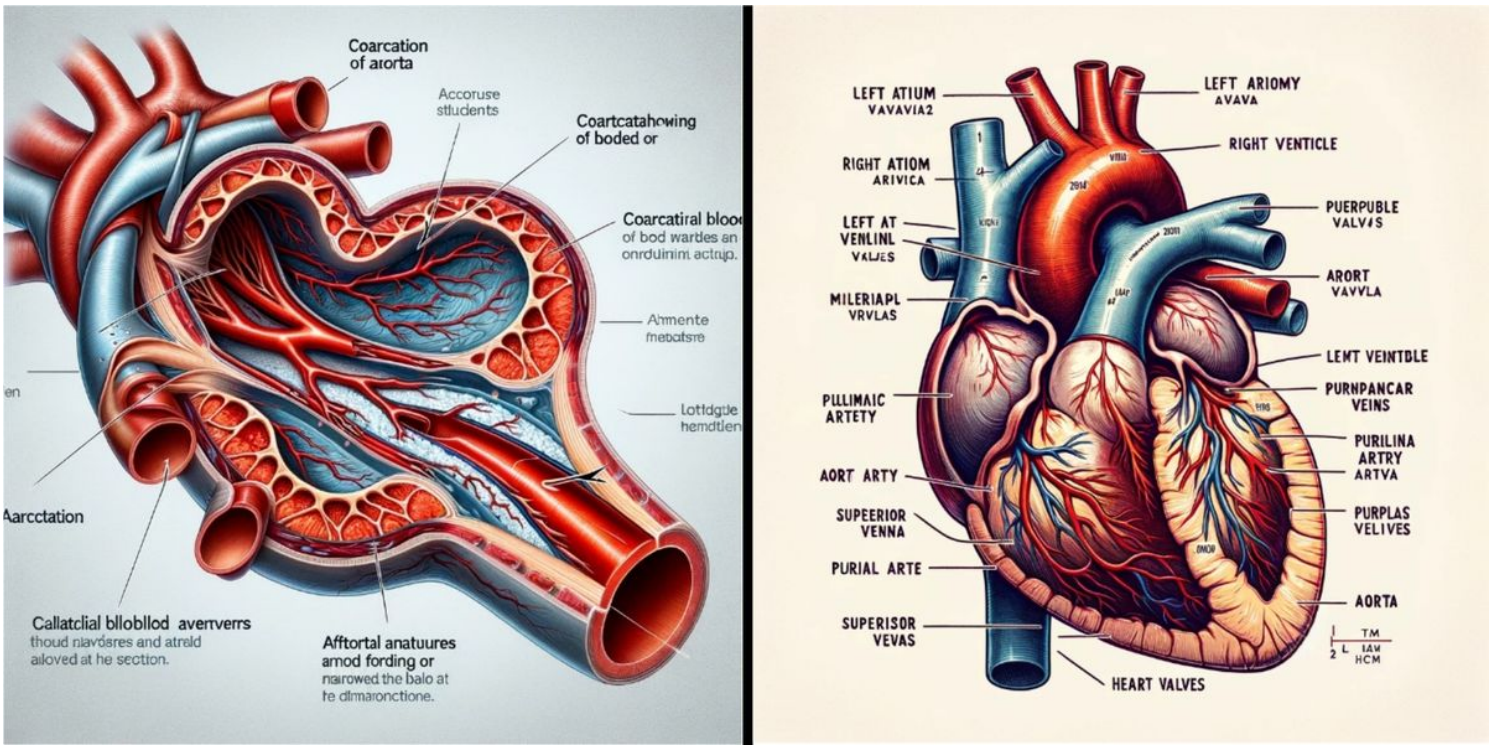

Figure 4

A: Example of the experts' rated "mostly fabricated" of the CHD image (coarctation of the aorta) as generated by DALL·E 3, Figure 4 B: Example of the experts' rated least fabricated rating of the "normal heart" as generated by DALL·E 3 (Compare to actual illustrations of coarctation of the aorta and normal heart in the CDC website: https://www.cdc.gov/ncbddd/heartdefects/coarctationofaorta.html accessed January 5th, 2024)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1.docx
- SupplementaryTable1.docx
- CHDDALLE.pdfCorrected.pdf