Journal of Clinical and Basic Research
Online ISSN: 2538-3736

Can AI illustrate science? A comparative benchmarking study of text-to-image artificial intelligence models for scientific communication

Reza Rakhshi [1], Teymoor Khosravi [1], Arian Rahimzadeh [1], Fatemeh Shahraki [1], Morteza Oladnabi [2,3]*

1. Student Research Committee, Golestan University of Medical Sciences, Gorgan, Iran
2. Gorgan Congenital Malformations Research Center, Golestan University of Medical Sciences, Gorgan, Iran
3. Ischemic Disorders Research Center, Golestan University of Medical Sciences, Gorgan, Iran

* Correspondence: Morteza Oladnabi. Gorgan Congenital Malformations Research Center, Golestan University of Medical Sciences, Gorgan, Iran. Tel: +9891044958390; Email: oladnabidozin@yahoo.com

Abstract

Background: This study investigates the capability of artificial intelligence (AI) to effectively illustrate scientific concepts for communication and educational purposes. It aims to evaluate the performance of three leading text-to-image Generative Artificial Intelligence (GenAI) models-Midjourney, DALL-E, and Stable Diffusion-in generating scientifically accurate visuals.

Methods: To assess the models, we employed a benchmarking approach, generating 120 images based on scientifically informed prompts. Each model's output was analyzed for aesthetics, core scientific concept representation, contextual relevance, and factual accuracy.

Results: The evaluation revealed that while GenAI models excelled in aesthetic quality, achieving a score of 90.83%, their success in capturing core scientific concepts was moderate at 48.30%. More concerning were the significant limitations in contextual relevance (9.17%) and factual accuracy, which scored a troubling 0%.

Conclusion: These findings underscore the current deficiencies of GenAI in producing effective educational illustrations. They highlight the urgent need for targeted training using domain-specific scientific datasets to enhance the precision of generated visual aids. Although the potential for AI in scientific communication is promising, substantial advancements are required to ensure both factual and contextual accuracy, facilitating a clearer understanding of complex concepts.

Article Type: Research Article

Article History

Received: 9 May 2025
Received in revised form: 5 June 2025
Accepted: 20 June 2025
Available online: 26 June 2025
DOI: 10.29252/JCBR.9.2.28

Keywords

Artificial intelligence
Computer graphics
Natural language comprehension

© The author(s)

Highlights

What is current knowledge?

- Generative AI models produce visually appealing images effectively.
- These models completely fail in achieving scientific accuracy.
- They show limited ability in maintaining contextual relevance.
- Scientific illustrations are vital for education and overcoming language barriers.
- Improving accuracy and contextual relevance is essential for scientific applications.

What is new here?

- GenAI needs targeted training for better educational illustration effectiveness.
- Substantial advancements are crucial for accurate scientific communication through AI.

Introduction

Artificial Intelligence (AI) is a field that focuses on creating intelligent systems that can mimic human intelligence, process data, and make autonomous decisions (1). Its goal is to develop machines that can solve problems and interact with the world (2). AI encompasses a diverse range of forms, each possessing distinct characteristics and serving specific purposes. These systems can be categorized into various subtypes, including machine learning, expert systems, natural language processing (NLP), planning systems, cognitive computing, robotics, and automation systems (3-5). Generative AI (GenAI) refers to a branch that focuses on creating systems capable of generating content, such as images, text, or music, autonomously. GenAI models utilize techniques such as deep learning and neural networks to learn patterns from existing data and generate new content that is coherent and resembles the training examples. These models have shown remarkable capabilities in tasks such as text-to-image synthesis, music composition, and even creating realistic human-like faces (6,7). Breakthroughs in the field of GenAI has been remarkable in recent years, with a growing application of GenAI in creative domains such as art, work, and research (8,9). Notably, text-to-image generation has gained significant popularity, exemplified by generative systems like Midjourney, Stable Diffusion, and DALL-E 3. These systems demonstrate the ability to synthesize images from textual prompts, often producing outputs that are virtually indistinguishable from those created by humans (10). Text generative AI models have demonstrated significant success in generating coherent and contextually relevant textual content across various domains and applications. For instance, we recently assessed the ability of OpenAI's ChatGPT to deliver answers to inquiries related to genetics (11). Based on our findings, the chatbot achieved a success rate of approximately 70 percent in providing accurate responses, which was comparable to that of human educated responders. However, when it comes to the realm of scientific figures and illustrations and visual teaching and learning methods, GenAI, specifically text-to-image generative AI models face significant challenges and often exhibit poor performance. Scientific figures and illustrations demand precision, accuracy, and a deep understanding of complex concepts, making them a unique domain for AI models. Scientific illustrations encompass a wide range of visual representations used to communicate scientific concepts, data, and observations (12). Charts/diagrams, infographics, and technical

[DOI: 10.66224/JCBR.9.2.28]

drawings are some common types. Assessing the effectiveness of these models is necessary to determine their efficacy and potential applications in scientific visualization and communication. Here in this study, we evaluated the performance of text-to-image generation AI models in drawing scientific illustrations and highlight their limitations and capabilities in visual teaching.

Methods

In this study, we conducted an evaluation of generated outputs from three GenAI models, namely Midjourney, DALL-E 3, and Stable Diffusion to assess their performance in scientific illustration and visual teaching .To ensure the scientific integrity and relevance of our evaluation, we developed a total of five sets, each consisting of 10 distinct prompts. These prompts were carefully engineered according to basic scientific knowledge. The prompt engeenering was guided by specific criteria, including the coverage of different scientific concepts, representation of various fields of science such as biology, mathematics, physics, and chemistry, and their direct relevance to the research question at hand. The evaluation criteria and metrics employed in this study were carefully chosen to provide a comprehensive assessment of the generated outputs. The first criterion, **conceptual capture**, assessed the AI models' proficiency in effectively reflecting the core idea or theme and its nuances presented in the prompts. **Contextual relatedness**, the second criterion, examined how related and pertinent the generated outputs were to the original prompts. **Factual accuracy**, the third criterion, scrutinized the outputs for their correctness and adherence to real-world truths as depicted in the prompts. The fourth criterion, **aesthetic visual quality**, evaluated the artistic merit and technical execution of the images generated by the models. Each response from the AI models was evaluated based on these four criteria, by a human workforce, also referred to as 'human labelers' in AI terminology, and categorized as either 'acceptable' or 'not acceptable'. By employing these four criteria, we aimed to obtain a holistic evaluation of the GenAI models' performance. This approach allowed us to assess outputs generated by the models, providing valuable insights into their capabilities and limitations.

Statistical analysis

To compare the performance of the three AI models (DALL-E, Midjourney, and Stable Diffusion) across the four evaluation criteria-Conceptual Capture, Contextual Relatedness, Factual Accuracy, and Aesthetic Visual Quality-we employed Chi-Square Tests of Independence. This test was selected due to the categorical nature of the evaluation data, which consisted of binary labels ("Acceptable" or "Not Acceptable") assigned by a human analytical panel to the images generated for 40 prompts across biology, mathematics, physics, and chemistry. The Chi-Square test assesses whether the distribution of these labels differs significantly across the three models for each criterion, testing the null hypothesis (H0) that there is no association between the model and the evaluation outcome.

For each criterion, contingency tables were constructed to summarize the frequency of "Acceptable" and "Not Acceptable" labels for each model. The Chi-Square statistic (χ^2), degrees of freedom (df), p-value, and effect size (Cramer's V) were calculated to evaluate the significance and magnitude of differences. Cramer's V was used as the effect size measure, with values interpreted as small ($V \approx 0.1$), medium ($V \approx 0.3$), or large ($V \approx 0.5$). The analyses were performed using Python (Version 3.9) with the "scipy.stats" library. For the Factual Accuracy criterion, where all models received "Not Acceptable" labels, statistical testing was not feasible due to the absence of variability, as the assumptions for Chi-Square testing (e.g., expected frequencies ≥ 5) were not met (13).

Results

A total of 120 outputs were generated by three distinct AI models-DALL-E v3, Midjourney v5.2, and Stable Diffusion v2.1-in response to a set of 40 unique prompts. These outputs, along with the prompts, are comprehensively documented in Supplementary Material 1. The evaluation process for the image outputs was based on four predetermined criteria: Conceptual Capture, Contextual Relatedness, Factual Accuracy, and Aesthetic Visual Quality. An analytical panel, consisting of human evaluators, examined each generated image and categorized them as 'acceptable' or 'not acceptable' based on their adherence to these criteria. Further detailed insights into the evaluation process are available in Supplementary Material 2, which contains tables detailing the labels assigned to each generated image by the human evaluators. Upon analyzing the performance of the individual models, it becomes evident that each model exhibits its own strengths and weaknesses across the evaluation criteria. According to the data summarized in Table 1, the overall performance of all three Generative AI models was found to be 48.30% for capturing the concept of the prompts, 9.17% in maintaining Contextual Relatedness, an observed 0% in Factual Accuracy, and 90.83% in achieving Aesthetic Visual Quality.

Figure 1 provides a comparative view of the performance of the three AI models across different criteria. The size of each slice in the pie charts directly corresponds to the model's strength in that particular criterion.

Table 1. The performance of current text-to-image generative artificial intelligence models in responding to prompts related to scientific illustration and visual teaching

Evaluation criteria	DALL-E v3 (n=40)	Midjourney v5.2 (n=40)	Stable Diffusion v2.1 (n=40)	Total Performance (n=120)
Conceptual capture	52.50%	40%	52.50%	48.30%
Contextual relatedness	10%	2.50%	15%	9.17%
Factual accuracy	0%	0%	0%	0%
Aesthetic visual quality	100%	95%	77.50%	90.83%

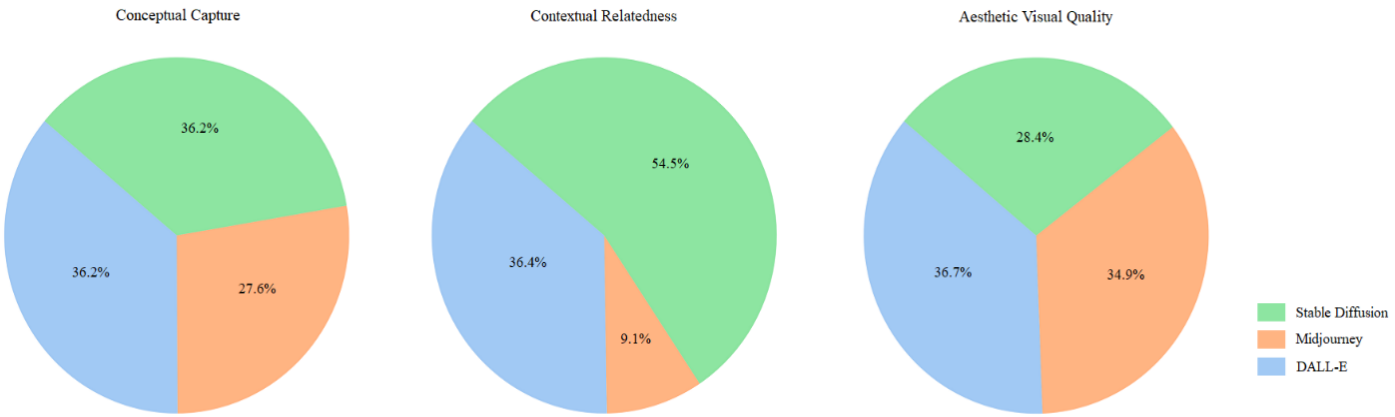

Figure 1. This figure presents three pie charts representing the average performance scores of three AI models-DALL-E v3, Midjourney v5.2, and Stable Diffusion v2.1-across three criteria: Conceptual Capture, Contextual Relatedness, and Aesthetic Visual Q

[DOI: 10.66224/JCBR.9.2.28]

For conceptual capture, the data shows that all three models have almost equal strength, with slight variations. Stable Diffusion and Midjourney both contribute around 36%, while DALL-E contributes around 28%. This suggests that all three models have similar capabilities in capturing the concept of the prompts. For contextual relatedness, Stable Diffusion dominates with over half of the pie chart attributed to it, indicating its superior performance in maintaining contextual relatedness. Midjourney follows at 36%, and DALL-E has the smallest slice at 9%. Moreover, for aesthetic visual quality, all models have comparable contributions to the chart, but Midjourney leads slightly at 37%, followed by DALL-E at 35%, and Stable Diffusion at 28%. This indicates that while all models generate aesthetically pleasing images, Midjourney and DALL-E do so slightly more consistently. It's crucial to mention that there is no pie chart for Factual Accuracy in Figure 2. This is because the performance of all models was 0% in this criterion, highlighting a significant area of improvement for all models. The proportional values behind this figure are summarized in Supplementary Material 3. These results highlight the strengths and weaknesses of each model. While all models demonstrated a strong ability to generate aesthetically pleasing images, they struggled to maintain contextual relatedness and factual accuracy. This suggests that while these models can generate visually appealing images, they may not accurately represent the factual content of the prompts, which is a crucial aspect to consider, especially in the context of scientific illustration and visual teaching.

Statistical analysis of model performance

To quantitatively compare the performance of the three AI models (DALL-E, Midjourney, and Stable Diffusion) across the four evaluation criteria-Conceptual Capture, Contextual Relatedness, Factual Accuracy, and Aesthetic Visual Quality-Chi-Square Tests of Independence were conducted for criteria with sufficient variability. The categorical nature of the evaluation labels ("Acceptable" vs. "Not Acceptable") necessitated the use of Chi-Square tests over ANOVA, which is more suitable for continuous data (14). For each criterion, contingency tables were constructed to summarize the distribution of labels across the models, and the Chi-Square statistic (χ^2), degrees of freedom (df), p-value, and effect size (Cramer's V) were calculated to assess the significance and magnitude of differences (13).

Conceptual capture

The contingency table for Conceptual Capture (Table 2) shows the frequency of "Acceptable" and "Not Acceptable" labels across the models.

Contextual relatedness

The contingency table for Contextual Relatedness (Table 3) summarizes the distribution of labels.

Aesthetic visual quality

The contingency table for Aesthetic Visual Quality (Table 4) is presented below.

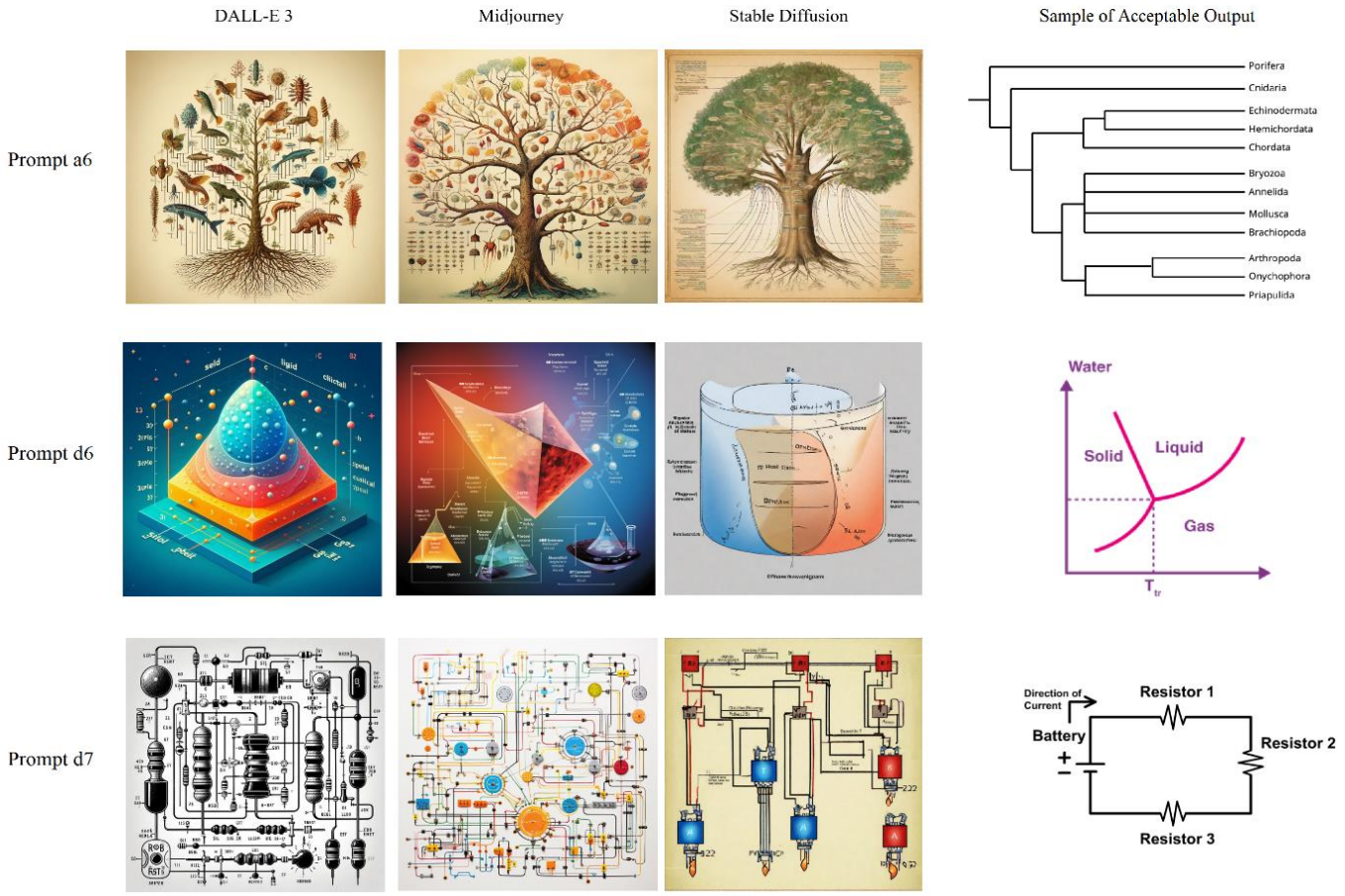

Figure 2. Comparative illustration of the results obtained from prompts a3, d6, and d7, along with an ideal output for each prompt.

Table 2. Contingency table for conceptual capture

Model	Acceptable	Not acceptable
DALL-E	18	22
Midjourney	12	28
Stable diffusion	18	22

The Chi-Square test revealed no significant difference in performance (χ^2 (2) = 2.857, p = 0.240, V = 0. Ditto)

[DOI: 10.66224/JCBR.9.2.28]

Factual accuracy

For Factual Accuracy, all models received "Not Acceptable" labels across all 40 prompts (Table 5), resulting in no variability and precluding statistical testing, as the assumptions for Chi-Square testing (e.g., expected frequencies ≥ 5) were not met (13).

Table 3. Contingency table for contextual relatedness

Model	Acceptable	Not acceptable
DALL-E	3	37
Midjourney	1	39
Stable diffusion	5	35

The Chi-Square test showed no significant difference (χ^2 (2) = 3.333, p = 0.189, V = 0.129), indicating similar performance in maintaining contextual relevance across the models

Table 4. Contingency table for contextual relatedness

Model	Acceptable	Not acceptable
DALL-E	40	0
Midjourney	37	3
Stable diffusion	34	6

The Chi-Square test indicated a significant difference (χ^2 (2) = 6.667, p = 0.036, V = 0.182), suggesting that the models differ in their ability to produce visually appealing outputs. The effect size (V = 0.182) indicates a small to medium difference, with DALL-E showing the highest proportion of "Acceptable" labels, followed by Midjourney and Stable Diffusion

Table 5. Contingency table for factual accuracy

Model	Acceptable	Not acceptable
DALL-E	0	40
Midjourney	0	40
Stable diffusion	0	40

These results highlight that while the models perform similarly in capturing concepts and contextual relevance, they differ significantly in aesthetic quality, with DALL-E achieving the highest performance in this criterion. The uniform failure in Factual Accuracy underscores a critical limitation of current general-purpose GenAI models for producing scientifically accurate illustrations, which is further discussed in the Limitations section

Discussion

Visualization-based teaching and learning methods, particularly through scientific illustrations, play a crucial role in enhancing understanding and knowledge retention in various disciplines (15,16). Scientific illustrations offer a powerful means of visually representing complex concepts, processes, and data in a comprehensible and engaging manner. By presenting information in a visual format, students are better able to grasp abstract ideas, visualize relationships between variables, and comprehend intricate scientific phenomena. One of the key advantages of scientific illustrations is their ability to simplify complex information. Through carefully designed diagrams, charts, and graphs, intricate scientific concepts can be distilled into visual representations that convey essential information concisely. This simplification aids learners in forming mental models, enabling them to grasp and remember complex scientific principles more effectively (17).

Moreover, scientific illustrations facilitate the communication of scientific ideas across different proficiency levels and language barriers (18). Visual representations can transcend linguistic limitations and provide a universal language for understanding scientific concepts. They allow individuals with diverse backgrounds and learning styles to access and interpret scientific information, fostering inclusivity and promoting scientific literacy (19). Furthermore, visualization-based teaching and learning methods encourage active participation and engagement among students. By incorporating interactive visual aids and tools, such as virtual simulations or augmented reality, learners can explore scientific phenomena, manipulate variables, and observe outcomes in a hands-on manner (20). This experiential learning approach enhances critical thinking skills, problem-solving abilities, and cultivates a deeper understanding of scientific principles. AI in education enhances learning through personalizing learning experiences, automating administrative tasks, and providing robust learning support systems. It facilitates individualized learning trajectories by adjusting to student feedback, while AI tools assist with grading and administrative tasks (21). Generative AI models also play a role, creating customized educational materials and interactive content, further enriching the learning experience. It aids educators in grading, assessing students, and identifying gaps in their understanding, hence allowing teachers to dedicate more time to teach and less on ancillary activities (6,22). In this study we observed that AI image generation models exhibit proficiency in certain domains while lacking in others. Despite the high rates of success in realizing aesthetic visual quality, the models' performance in areas such as contextual relatedness and factual accuracy was underwhelming. This discrepancy in performance capacity suggests that the visual appeal of the outputs does not necessarily translate to their utility in conveyance of precise and contextually pertinent information. The models, including Stable Diffusion, Midjourney, and DALL-E, showcased a combined efficiency for conceptual capture, suggesting that while they can grasp the abstract or thematic essence of user prompts, the nuanced and factual details required for accurate representation remain elusive. This is further underlined by the significant gap observed in factual accuracy across all models, which was starkly highlighted at 0%. Figure 2 presents the results for prompts a3, d6, and d7, accompanied by a straightforward illustration of a suitable output for each. These instances exemplify the difficulties GenAIs tools encounter when attempting to grasp the multifaceted nature of scientific concepts. The figure underscores the observed shortcomings in factual and contextual representations produced by the GenAIs.

The images produced by these models tend to prioritize artistic aesthetics rather than adhering strictly to scientific representation. The lack of informativeness in the generated images implies that key details and essential information required for a comprehensive understanding of the schematic figure and flowchart prompts are either missing or inadequately conveyed. This can hinder the effective communication of scientific concepts or data. Furthermore, the observed lack of accuracy raises concerns regarding the fidelity of the AI generated images to the intended scientific content.

Limitation

Additionally, the reliance on general-purpose GenAI models may limit the precision of outputs for highly specialized scientific illustrations. Notably, for the Factual Accuracy criterion, all models received "Not Acceptable" labels across all 40 prompts, indicating no variability and precluding statistical comparisons using tests like Chi-Square, as the assumptions for such tests (e.g., expected frequencies ≥ 5) were not met (13). This uniform failure highlights a significant challenge in using current GenAI models for scientifically accurate illustrations, suggesting a need for further development in this area. To address these limitations, a potential solution lies in the development of domain-specific GenAI (DSGAI) models (23,24) tailored for creating scientific illustrations. By training DSGAIs on vast datasets comprising scientific diagrams, charts, and illustrations from a wide range of disciplines, these models can 'learn' the intricacies of scientific representation, ensuring a higher level of fidelity and precision in their generated outputs. The implementation of DSGAIs holds promise for researchers, educators, and communicators across various scientific fields, aiding in the creation of visually compelling and scientifically rigorous illustrations for knowledge dissemination, research presentation, and educational materials.

Future work

To enhance the generalizability of our findings, future research could expand the prompt set to include a broader range of scientific disciplines, such as social sciences, environmental sciences, and humanities, potentially increasing the number of prompts to 100 or more. This would allow for a more comprehensive evaluation of [Your technology/method, e.g., GenAI for scientific illustrations] across diverse domains. Additionally, the development and evaluation of domain-specific GenAI (DSGAI) models trained on diverse datasets could further improve the accuracy and applicability of generated scientific illustrations, addressing the limitations of general-purpose models and supporting a wider range of scientific communication needs.

[DOI: 10.66224/JCBR.9.2.28]

Ethical implications

The use of AI in educational settings can inadvertently propagate misinformation if outputs are not rigorously validated. As Selwyn (2022) notes, AI-driven educational tools risk disseminating inaccurate or oversimplified content, especially in rapidly evolving fields such as [Insert relevant field, e.g., medical education]. This can erode trust and mislead learners. To mitigate this, we recommend cross-referencing system outputs with peer-reviewed sources and involving domain experts in content curation to ensure accuracy and reliability (25).

Training datasets often reflect systemic biases, such as the over-representation of certain scientific fields (e.g., physics or computer science) due to their prominence in available data (26). This can lead to skewed outputs that marginalize underrepresented disciplines, such as social sciences or humanities, perpetuating an incomplete view of scientific inquiry (27). For instance, a model trained on datasets with limited qualitative research may undervalue its contributions. To address this, we propose diversifying datasets to include a broader range of scientific perspectives and conducting regular audits of training data to identify and correct biases (26).

To navigate these ethical challenges, transparency in algorithmic processes and interdisciplinary collaboration are essential (28). Engaging stakeholders, such as educators and researchers, can help identify real-world implications and ensure equitable outcomes. Continuous monitoring of AI will also be critical to adapting to evolving ethical standards.

Conclusion

The evaluation of DALL-E v3, Midjourney v5.2, and Stable Diffusion v2.1 reveals that while these generative AI models excel in producing aesthetically pleasing images, with an overall Aesthetic Visual Quality score of 90.83%, they fall short in Contextual Relatedness (9.17%) and Factual Accuracy (0%). This indicates a critical limitation in their ability to generate scientifically accurate and contextually relevant illustrations, which is essential for effective visual teaching and scientific communication. These findings suggest that while current text-to-image AI models are promising for creating visually engaging content, significant improvements are needed to enhance their precision and reliability for educational and scientific applications.

Acknowledgement

We are grateful for the support provided by the Medical Genetics Department of Golestan University of Medical Sciences.

Funding sources

This work was supported by the Golestan University of Medical Sciences and Health Services (Grant Number: 113845)

Ethical statement

This study was ethically approved by Ethics Committee of Golestan University of Medical Sciences (Ethics Code: IR.GOUMS.REC.1402.458)

Conflicts of interest

No potential conflict of interest was reported by the author(s).

Author contributions

Conceptualization: Reza Rakhshi, Teymoor Khosravi; Data Curation: Teymoor Khosravi, Fatemeh Shahraki, Arian Rahimzadeh; Formal Analysis: Arian Rahimzadeh, Fatemeh Shahraki; Funding Acquisition: Morteza Oladnabi, Teymoor Khosravi; Investigation: Arian Rahimzadeh; Methodology: Reza Rakhshi; Project Administration: Morteza Oladnabi; Resources: Teymoor Khosravi; Software: Teymoor Khosravi, Arian Rahimzadeh; Supervision: Morteza Oladnabi; Validation: Reza Rakhshi, Fatemeh Shahraki; Visualization: Reza Rakhshi, Teymoor Khosravi Arian Rahimzadeh, Fatemeh Shahraki; Writing -Original Draft: Reza Rakhshi, Teymoor Khosravi, Arian Rahimzadeh; Writing -Review &Editing: Reza Rakhshi, Teymoor Khosravi Morteza Oladnabi.

Data availability statement

No additional data were created or used in this study beyond what is presented in the manuscript

References

1. Ertel W. Machine Learning and Data Mining. Introduction to artificial intelligence: Springer; 2018;175-243. [View at Publisher] [DOI] [Google Scholar]
2. Martinez R. Artificial intelligence: Distinguishing between types & definitions. Nevada Law Journal. 2019;19(3):9. [View at Publisher] [Google Scholar]
3. Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, et al. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. Mathematics. 2022;10(15):2552. [View at Publisher] [DOI] [Google Scholar]
4. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: An artificial intelligence approach: Springer Science and Business Media; 2013. [View at Publisher] [Google Scholar]
5. Goldberg Y. Neural network methods for natural language processing: Springer Nature; 2022. [View at Publisher] [Google Scholar]
6. Baidoo-Anu D, Ansah LO. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI. 2023;7(1):52-62. [View at Publisher] [DOI] [Google Scholar]
7. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science.2023;381(6654):187-92. [View at Publisher] [DOI] [Google Scholar]
8. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. "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. 2023;71:102642. [View at Publisher] [DOI] [Google Scholar]
9. Brynjolfsson E, Li D, Raymond LR. Generative AI at work. National Bureau of Economic Research; 2023. [View at Publisher] [DOI] [Google Scholar]
10. Enjellina ,Beyan EVP, Rossy AGC. A Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field. Journal of Artificial Intelligence in Architecture. 2023;2(1):53-65. [View at Publisher] [DOI] [Google Scholar]
11. Khosravi T, Al Sudani ZM, Oladnabi M. To what extent does ChatGPT understand genetics? Innovations in Education and Teaching International. 2023:1320-9. [View at Publisher] [DOI] [Google Scholar]
12. Richards C. Technical and scientific illustration. Information design: Research and practice. 2017:85-106. [View at Publisher] [Google Scholar]
13. Kim H-Y. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod. 2017;42(2):152-5. [View at Publisher] [DOI] [Google Scholar]
14. Gaddis G. Advanced biostatistics: Chi-square, ANOVA, regression, and multiple regression. Doing Research in Emergency and Acute Care: Making Order Out of Chaos. 2015:213-22. [View at Publisher] [DOI] [Google Scholar]
15. Kara S. Prospective Visual Arts Teachers' Innovation Skills and Attitudes towards Computer Assisted Instruction. International Journal of Technology in Education and Science. 2020;4(2):98-107. [View at Publisher] [DOI] [Google Scholar]
16. Serafini F. Reading the visual: An introduction to teaching multimodal literacy: Teachers College Press; 2014. [View at Publisher] [Google Scholar]
17. Hafeez M. Systematic review on modern learning approaches, critical thinking skills and students learning outcomes. Indonesian Journal Of Educational Research and Review. 2021;4(1):167-78. [View at Publisher] [DOI] [Google Scholar]
18. Amano T, González-Varo JP, Sutherland WJ. Languages are still a major barrier to global science. PLoS biology. 2016;14(12):e2000933. [View at Publisher] [DOI] [PMID] [Google Scholar]
19. Amano T, Rios Rojas C, Boum II Y, Calvo M, Misra BB. Ten tips for overcoming language barriers in science. Nat Hum Behav. 2021;5(9):1119-22. [View at Publisher] [DOI] [PMID] [Google Scholar]

[DOI: 10.66224/JCBR.9.2.28]

20. Wang M, Wu B, Kinshuk, Chen N-S, Spector JM. Connecting problem-solving and knowledge-construction processes in a visualization-based learning environment. Comput Educ. 2013;68:293-306. [View at Publisher] [DOI] [Google Scholar]
21. Zhang K, Aslan AB. AI technologies for education: Recent research & future directions. Comput Educ.: Artif Intell. 2021;2:100025. [View at Publisher] [DOI] [Google Scholar]
22. Lim WM, Gunasekara A, Pallant JL, Pallant JI, Pechenkina E. Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The International Journal of Management Education. 2023;21(2):100790. [View at Publisher] [DOI] [Google Scholar]
23. Liang P. Semi-supervised learning for natural language: Massachusetts Institute of Technology; 2005. [View at Publisher] [Google Scholar]
24. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018. [View at Publisher] [Google Scholar]
25. Selwyn N. The future of AI and education: Some cautionary notes. Eur J Educ. 2022;57(4):620-31. [View at Publisher] [DOI] [Google Scholar]
26. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021;610-23. [View at Publisher] [DOI] [Google Scholar]
27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. [View at Publisher] [DOI] [PMID] [Google Scholar]
28. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds Mach. 2020;30(4):681-94. [View at Publisher] [DOI] [Google Scholar]

How to Cite:
Rakhshi R, Khosravi T, Rahimzadeh A, Shahraki F, Oladnabi M. Can AI illustrate science? A comparative benchmarking study of text-to-image artificial intelligence models for scientific communication. *JCBR.* 2025;9(2):28-33. http://dx.doi.org/10.29252/JCBR.9.2.28

[DOI: 10.66224/JCBR.9.2.28]