

Research Article

Exploring the Boundaries of AI: ChatGPT's Accuracy in Anatomical Image Generation & Bone Identification of the Foot

Simran Shamith^{1a}, Neshal K. Kothari^{1b}, Serena K. Kothari^{1c}, Carolyn Giordano, Ph.D.^{1d}

¹ Drexel University

Keywords: ChatGPT, Medical Education, Orthopedic Anatomy

<https://doi.org/10.60118/001c.128540>

Journal of Orthopaedic Experience & Innovation

Vol. 6, Issue 1, 2025

Introduction

Large language models like ChatGPT are used in fields ranging from education to healthcare. This technology can simplify complex topics and help students study topics like anatomy. However, many concerns exist regarding its accuracy. This study aims to evaluate the image recognition and creation of DALL-E, ChatGPT's native image generation engine, by examining its ability to label bones of the human foot and generate an accurate anatomical illustration.

Methods

This study utilized an untrained version of ChatGPT (version 4o). Two tasks were performed. First, ChatGPT was prompted to generate a labeled anatomical image of the human foot. The generated image was compared against anatomical references to assess its accuracy. In the second task, the model was asked to identify bones in three unlabeled foot images, two illustrations and one x-ray. Each image was inputted, with ChatGPT's

-
- a Simran Shamith is a third-year medical student at Drexel University College of Medicine with a focus in academic medicine. Her past research explores survey validation using ChatGPT, the impact of co-morbidities on outcomes in orthopaedic surgery, and reducing opioid use following surgery. With a strong interest in AI in medicine, medical education, and orthopaedic surgery, Simran is dedicated to advancing healthcare through innovative research and technology.

[Visit Simran Shamith](#)

[Conflicts of Interest Statement for Simran Shamith](#)

- b Neshal Kothari is a second-year physics student at Drexel University's College of Arts and Sciences in Philadelphia, PA. His research focuses on computational lipid behavior and the application of AI tools in medicine. He holds a leadership position in the Society of Physics Students, and volunteers in educational events to teach physics to younger students.

[Connect with Neshal Kothari on LinkedIn](#)

[Conflicts of Interest Statement for Neshal Kothari](#)

- c Serena K. Kothari is currently a second year medical student with a heavy interest in orthopaedic surgery. She enjoys exploring various areas within orthopedics including medical education.

[Connect with Serena Kothari on LinkedIn](#)

[Conflicts of Interest Statement for Serena Kothari](#)

- d Dr. Carolyn Giordano is the associate dean for assessment and evaluation and an associate professor in the Department of Family, Community and Preventive Medicine at Drexel College of Medicine. She has been the primary author or co-author on over a dozen articles in peer-reviewed journals, in the area of health and medical education as well as in interprofessional education, and is the author and co-author on two book chapters on assessment and data analytics. Giordano is a distinguished psychology fellow in the National Academy of Practice as well as a fellow in the Association of Schools Advancing Health Professions where she is a member of the interprofessional education subcommittee and has co-edited a special edition of the Journal of Allied Health dedicated to interprofessional education works. She has been invited to give testimony to HRSA on the impact of interprofessional education and has also been awarded the distinguished paper award from the Association of Schools of Allied Health Professions and the James B. Erdmann, PhD Award for Excellence in Interprofessional Education & Collaborative Practice.

[Visit Dr. Carolyn Giordano](#)

[Connect with Dr. Carolyn Giordano on LinkedIn](#)

[Conflicts of Interest Statement for Dr. Carolyn Giordano](#)

memory cleared between each prompt to prevent learning. The responses were compared to referenced anatomical labels, and accuracy was calculated for each.

Results

The study revealed inaccuracies in ChatGPT's ability to generate accurate anatomical images and to identify parts of an uploaded image. While ChatGPT produced a visually detailed drawing, several labels were misplaced, critical bones were incorrect or not included, and entire structures were missing from the image. In the second labeling task, accuracy was low, with correct labeling rates of 27%, 57%, and 0% for each of the three image prompts. ChatGPT frequently misidentified and mislabeled bones.

Discussion

The findings highlight ChatGPT's current limitations as an educational tool, particularly in its untrained state. The inaccuracies found underscore the challenges of using language-based AI models for specialized fields like anatomy. Factors contributing to these errors include ChatGPT's reliance on text-based training data, which lacks the depth required for accurate visual representation. In addition to inaccurate data, the errors may also be due to ChatGPT not inherently having spatial or depth awareness – it understands patterns in data but not the geometry of real-world objects or figures. This study suggests that while AI holds promise as a supplementary educational resource, users must remain vigilant and always verify AI-generated content.

This article, like all JOEI articles, is available for continuing education

Reflect Now!

Click here : <https://joeipub.com/learning>

INTRODUCTION

Chat generative pre-trained transformer (ChatGPT) is an artificial intelligence (AI) tool developed by OpenAI to generate human-like text, allowing it to engage in conversations, answer complex questions, and assist with tasks. It is a large language model that is used across healthcare and education for coding, data analysis, writing, and more (Dave, Athaluri, and Singh 2023; Mir et al. 2023). These models are rapidly evolving and can recognize, summarize, translate, predict and generate not only written content, but photographs, audio and video files as well (Eysenbach 2023).

ChatGPT is growing exponentially with over 200 million weekly users globally (Babu 2024). Among its many users, ChatGPT can benefit individuals looking for medical information (Khlaif et al. 2023; Xu, Chen, and Miao 2024). For medical students, AI is a valuable tool to quickly master a breadth of concepts, including complex anatomic relationships. Medical students are leveraging AI to organize lectures, create efficient study schedules, and understand unusual disease presentations (Tolentino et al. 2024). Prior studies have shown that ChatGPT can perform at a level comparable to a medical student, even passing medical board examinations (Ghanem et al. 2023). Educators are learning how to incorporate AI into their workflow and to teach students how to use AI in an effective and ethical

manner (Mir et al. 2023). Past studies highlight ChatGPT's future potential as an educational resource, especially in medicine, however, some are concerned about its validity and reliability (Ray 2023; Alkaissi and McFarlane 2023).

Beyond student education, patients are turning to AI to initiate queries on their conditions and to gain a better understanding of their diagnoses and information given by their providers. ChatGPT translates medical terminology into simple everyday language. For example, a patient who has a fracture might use ChatGPT to explore their injury, understand the healing process, and learn about potential treatments. However, studies have shown inaccuracies in the answers provided by ChatGPT regarding medical questions posed by patients (Jagiella-Lodise, Suh, and Zelenski 2024). A systematic review in 2023 showed that while ChatGPT may help answer patient questions, there are problems with accuracy and bias in the answers ChatGPT provides (Garg et al. 2023).

In April of 2022, ChatGPT expanded to include AI image generations with the tool DALL-E. The DALL-E engine processes sequences of text and interprets them as individual tokens to create visual representations (Ramesh et al. 2021; "Dall-E: Creating Images from Text" 2021). This creates the potential for AI models to illustrate physiological processes and anatomical relationships of the human body. These images can be used as study tools for medical students or in patient education to create visual aids, espe-

cially relevant in the field of orthopedic surgery. However, like ChatGPT's earlier written output, the accuracy of the information and images produced by AI models can vary and may lack the depth of understanding of a trained professional (Leivada, Murphy, and Marcus 2023). Past studies have shown ChatGPT's inaccuracy in generating scientific writing, often including a mix of accurate and fabricated data (Dave, Athaluri, and Singh 2023).

With the advent of image generation, it is important to assess the accuracy of anatomical images created by ChatGPT and DALL-E. A study by Adams et al. showed DALL-E's ability to generate x-ray images of normal anatomy including the skull, chest, pelvis, hand, and foot. However, the study demonstrated several inaccuracies when ChatGPT was asked to generate MRI, CT, or ultrasound images. In addition, when asked to generate an erased portion of an x-ray, there were often large errors when x-rays crossed joint lines such the shoulder or the tarsal bones. Adams et al. did not explore ChatGPT's ability to label the radiographic images or generate illustrations of human anatomy (Adams et al. 2023).

This study aims to explore the accuracy of ChatGPT version 4o, without any specialized training in anatomy, in identifying bone anatomy and generating an anatomical illustration of the human foot. By assessing the model's performance in these tasks, we aim to evaluate ChatGPT and DALL-E as a tool for medical and patient education in human anatomy, a topic many find difficult to learn. Studies have shown students struggle with anatomy due to the depth of knowledge required and difficulty in visualizing anatomic structures and relationships (Cheung, Bridges, and Tipoe 2021).

METHODS

This study was divided into two aims, anatomical image generation and structure identification on a widely recognizable structure, the human foot. Two prompts were entered into ChatGPT version 4o on September 27th, 2024. ChatGPT's memory was cleared in between each prompt to remove the chance of learning from the previous answer. Only the untrained model of ChatGPT was tested. The responses from the models were recorded and compared to the correct anatomy. The number of incorrect responses was noted for each image. An average percentage of correct and incorrect responses was calculated for each version of ChatGPT.

ANATOMICAL IMAGE GENERATION

For the anatomical image generation task, ChatGPT was asked in layman's terms to produce an image of the human foot with the bones labeled using the following prompt.

Prompt 1: 'Make me a picture of human foot with the bones labeled'

The generated image was qualitatively compared to standard anatomical references to assess the accuracy of bone representation and labeling.

STRUCTURE IDENTIFICATION

Two unlabeled illustrated images, figures 1 and 2, of the foot and one unlabeled x-ray image, figure 3, were selected. Each image was inputted into ChatGPT Version 4o and the model was prompted to identify the bones depicted in each image using the following prompt.

Prompt 2: 'Identify the structures labeled in this image'

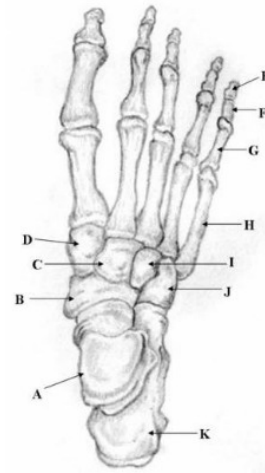


Figure 1.

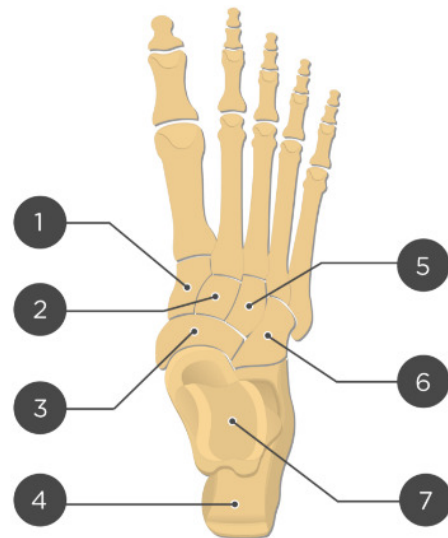


Figure 2.



Figure 3.

RESULTS

Two prompts were entered into ChatGPT (Prompt 1: "Make me a picture of human foot with the bones labeled", Prompt 2: "Identify the structures labeled in this image"). Results of prompt 1, seen in figure 4, are that the generated image was detailed and visually appealing, correctly displaying a foot with visible bones as requested. However, the labeling is entirely inaccurate, with misspelled anatomical terms, misplaced labels, and several missing bones, including the phalanges of the 4th toe. The result of prompt 2 for the first illustrated image is 27% accurate with 3 of 11 bones accurately identified as shown in table 1. The result of prompt 2 for the second illustrated image is 57% accurate with 4 of 7 bones accurately identified as shown in table 2. The result of prompt 2 for the x-ray image is 0% accurate with 0 of 19 structures accurately identified as shown in table 3.

Results are shown below.

Prompt 1: "Make me a picture of human foot with the bones labeled"

ChatGPT Output:

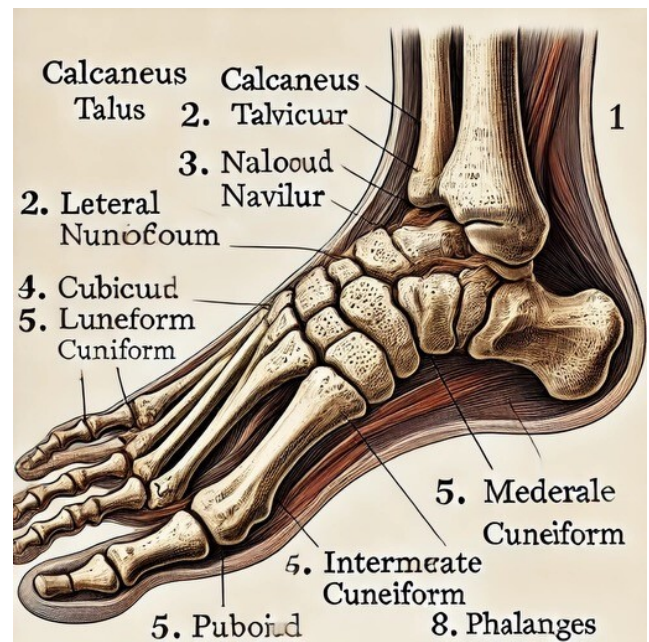


Figure 4.

Prompt 2: Identify the structures labeled in this image

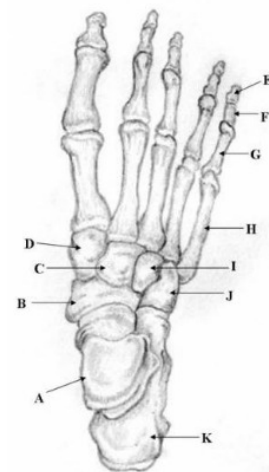


Figure 1.

Out of 11 labels, 3 were identified correctly, resulting in an accuracy of 27%.

Prompt 2: Identify the structures labeled in this image

Table 1. Output of ChatGPT for figure 1 anatomical labeling

ChatGPT Output	Anatomical structure	Correct or Incorrect
A: Calcaneus (heel bone)	A: Talus	Incorrect
B: Talus	B: Navicular	Incorrect
C: Navicular	C: Intermediate cuneiform	Incorrect
D: Medial cuneiform	D: Medial cuneiform	Correct
E: Distal phalanx of the first toe	E: Distal phalanx of the fifth toe	Incorrect
F: Distal phalanx of the second toe	F: Middle phalanx of the fifth toe	Incorrect
G: Distal phalanx of the third toe	G: Proximal phalanx of the fifth toe	Incorrect
H: Intermediate cuneiform	H: Metatarsal	Incorrect
I: Lateral cuneiform	I: Lateral cuneiform	Correct
J: Cuboid	J: Cuboid	Correct
K: Tuberosity of the fifth metatarsal	K: Calcaneus	Incorrect

Table 2. Output of ChatGPT for figure 2 anatomical labeling

ChatGPT Output	Anatomical structure	Correct or Incorrect
1. Navicular bone	1. Medial cuneiform	Incorrect
2. Intermediate cuneiform	2. Intermediate cuneiform	Correct
3. Talus	3. Navicular	Incorrect
4. Calcaneus	4. Calcaneus	Correct
5. Lateral cuneiform	5. Lateral cuneiform	Correct
6. Cuboid bone	6. Cuboid	Correct
7. Medial cuneiform	7. Talus	Incorrect

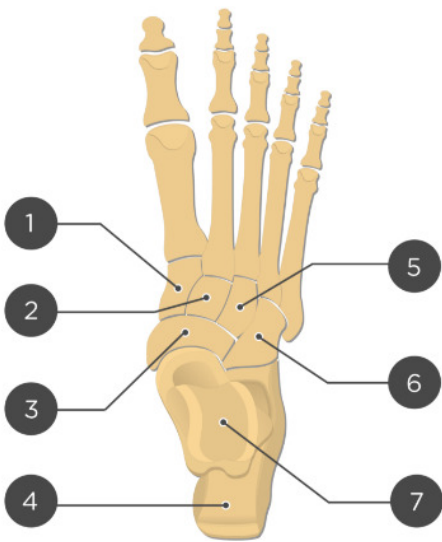


Figure 2.

Out of 7 labels, 4 were identified correctly, resulting in an accuracy of 57%.

Prompt 2: Identify the structures labeled in this image



Figure 3.

Out of 19 labels, 19 were identified correctly, resulting in an accuracy of 0%.

Table 3. Output of ChatGPT for figure 3 anatomical labeling

ChatGPT Output	Anatomical Structure	Correct or Incorrect
1. Calcaneus (heel bone)	1. Medial malleolus	Incorrect
2. Talus (ankle bone)	2. Lateral malleolus	Incorrect
3. Navicular	3. Head of talus	Incorrect
4. Cuboid	4. Calcaneus	Incorrect
5. Medial cuneiform	5. Navicular	Incorrect
6. Intermediate cuneiform	6. Cuboid	Incorrect
7. Lateral cuneiform	7. Base of fifth metatarsal	Incorrect
8. Base of the 1st metatarsal	8. Lateral cuneiform	Incorrect
9. Base of the 2nd metatarsal	9. Intermediate cuneiform	Incorrect
10. Base of the 3rd metatarsal	10. Medial cuneiform	Incorrect
11. Proximal phalanx of the hallux	11. Medial sesamoid	Incorrect
12. Proximal phalanx of the 2nd toe	12. Lateral sesamoid	Incorrect
13. Proximal phalanx of the 3rd toe	13. Shaft of second metatarsal	Incorrect
14. Proximal phalanx of the 4th toe	14. Neck of third metatarsal	Incorrect
15. Proximal phalanx of the 5th toe	15. Head of fourth metatarsal	Incorrect
16. Middle phalanx of the 2nd toe	16. Metatarsophalangeal joint	Incorrect
17. Middle phalanx of the 3rd toe	17. Proximal phalanx of 3 rd toe	Incorrect
18. Middle phalanx of the 4th toe	18. Middle phalanx of 2 nd toe	Incorrect
19. Middle phalanx of the 5th toe	19. Distal phalanx of 2 nd toe	Incorrect

DISCUSSION

The output from ChatGPT v4o of prompt 1 "*Make me a picture of human foot with the bones labeled*" demonstrates a visually appealing and detailed depiction of a human foot with visible bones. However, the labeling is inaccurate, with anatomical terms misspelled, labels misplaced, and several key bones, such as the phalanges of the 4th toe, completely missing. Prompt 2 reveals varying accuracy between the 3 images used: 27% for the first illustrated image, 57% for the second illustrated image, and 0% accuracy for the x-ray image.

These findings indicate that currently, as of September 2024, ChatGPT version 4o, utilizing the image generator DALL-E, cannot accurately identify uploaded images or generate accurate images, particularly in the context of anatomical study. Like the process of image generation, ChatGPT breaks down a picture into several visual tokens and creates "chunks" of data. The program then looks for patterns based on its previous data. Depending on the accuracy or level of detail in the previous data, ChatGPT may misinterpret or generalize an input (Ramesh et al. 2021).

Several other factors may have also contributed to these inaccuracies. First, it's important to understand that ChatGPT operates primarily as a language model. It generates responses based on extensive datasets rather than engaging in complex thoughts as humans do. This means that, while the responses can *sound* reasonable, they may lack verification or more detailed thought (Bisk et al. 2020). ChatGPT's training dataset may include a wide array of visual content which may also result in biased interpretations instead of factually accurate representations (Ray 2023).

Another challenge is the difficulty of translating complex text descriptions into visual representations. Users may ask for detailed images, but ChatGPT can misinterpret or oversimplify these requests depending on the past data it has available. This can lead to generated images that significantly differ from what the user intended. Additionally, if the model is trained on limited or inaccurate information about a specific topic, it will naturally produce flawed outputs (Dave, Athaluri, and Singh 2023; "Dall-E: Creating Images from Text" 2021; Huang, Wang, and Yang 2023).

The implications of these inaccuracies are particularly worrisome for students and professionals in healthcare. While medical students and professionals are generally equipped to sift through information critically, patients or caregivers with little medical knowledge and low health literacy may not have the skills to distinguish between accurate and inaccurate information. As past studies have also demonstrated, this study highlights the importance of users approaching AI-generated content with skepticism and verifying information from other reliable sources (Alkaissi and McFarlane 2023; Jagiella-Lodise, Suh, and Zelenksi 2024; Leivada, Murphy, and Marcus 2023).

In conclusion, while tools like ChatGPT hold great potential for changing how we access information, it is crucial for users to remain vigilant in verifying any important data gathered from it. Improving accuracy is a shared responsibility between developers and users, who need to apply critical thinking when accessing AI-generated content.

Submitted: November 09, 2024 EDT. Accepted: January 18, 2025 EDT. Published: July 03, 2025 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-NC-ND-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-nc-nd/4.0> and legal code at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> for more information.

REFERENCES

- Adams, L. C., F. Busch, D. Truhn, M. R. Makowski, H. J. W. L. Aerts, and K. K. Bressen. 2023. "What Does DALL-E 2 Know About Radiology?" *Journal of Medical Internet Research* 25:e43110. <https://doi.org/10.2196/43110>.
- Alkaissi, H., and S. I. McFarlane. 2023. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing." *Cureus* 15 (2): e35179. <https://doi.org/10.7759/cureus.35179>.
- Babu, J. 2024. "OpenAI Says Chatgpt's Weekly Users Have Grown to 200 Million." Reuters. August 29, 2024. <https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/>.
- Bisk, Y., A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, et al. 2020. "Experience Grounds Language." *arXiv.Org*, November. <https://doi.org/10.18653/v1/2020.emnlp-main.703>.
- Cheung, C. C., S. M. Bridges, and G. L. Tipoe. 2021. "Why Is Anatomy Difficult to Learn? The Implications for Undergraduate Medical Curricula." *Anatomical Sciences Education* 14 (6): 752–63. <https://doi.org/10.1002/ase.2071>.
- "Dall-E: Creating Images from Text." 2021. OpenAI. January 5, 2021. <https://openai.com/index/dall-e/>.
- Dave, T., S. A. Athaluri, and S. Singh. 2023. "ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations." *Frontiers in Artificial Intelligence* 6:1169595. <https://doi.org/10.3389/frai.2023.1169595>.
- Eysenbach, G. 2023. "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers." *JMIR Med Educ* 9:e46885. <https://doi.org/10.2196/46885>.
- Garg, R. K., V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, and S. K. Kar. 2023. "Exploring the Role of ChatGPT in Patient Care (Diagnosis and Treatment) and Medical Research: A Systematic Review." *Health Promotion Perspectives* 13 (3): 183–91. <https://doi.org/10.34172/hpp.2023.22>.
- Ghanem, D., O. Covarrubias, M. Raad, D. LaPorte, and B. Shafiq. 2023. "ChatGPT Performs at the Level of a Third-Year Orthopaedic Surgery Resident on the Orthopaedic In-Training Examination." *JB & JS Open Access* 8 (4): e23.00103. <https://doi.org/10.2106/JBJS.OA.23.00103>.
- Huang, Z., M. Wang, and K. Yang. 2023. "The Impact of CHATGPT on the Learning Efficacy of Chinese College Students." EWA Publishing. November 20, 2023. <https://www.ewadirect.com/proceedings/Inep/article/view/6590>.
- Jagiella-Lodise, O., N. Suh, and N. A. Zelenski. 2024. "Can Patients Rely on ChatGPT to Answer Hand Pathology-Related Medical Questions?" *Hand (New York, N.Y.)* 15589447241247246. <https://doi.org/10.1177/15589447241247246>.
- Khlaif, Z. N., A. Mousa, M. K. Hattab, J. Itmazi, A. A. Hassan, M. Sanmugam, and A. Ayyoub. 2023. "The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation." *JMIR Medical Education* 9:e47049. <https://doi.org/10.2196/47049>.
- Leivada, E., E. Murphy, and G. Marcus. 2023. "DALL-E 2 Fails to Reliably Capture Common Syntactic Processes." *Social Sciences & Humanities Open* 8 (1): 100648. <https://doi.org/10.1016/j.ssaho.2023.100648>.
- Mir, M. M., G. M. Mir, N. T. Raina, S. M. Mir, S. M. Mir, E. Miskeen, M. H. Alharthi, and M. M. S. Alamri. 2023. "Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives." *Journal of Advances in Medical Education & Professionalism* 11 (3): 133–40. <https://doi.org/10.30476/JAMP.2023.98655.1803>.
- Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. 2021. "Zero-Shot Text-to-Image Generation." *ArXiv*. <https://doi.org/10.48550/arXiv.2102.12092>.
- Ray, P. P. 2023. "CHATGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." *Internet of Things and Cyber-Physical Systems*, April. <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Tolentino, R., A. Baradaran, G. Gore, P. Pluye, and S. Abbasgholizadeh-Rahimi. 2024. "Curriculum Frameworks and Educational Programs in AI for Medical Students, Residents, and Practicing Physicians: Scoping Review." *JMIR Medical Education* 10:e54793. <https://doi.org/10.2196/54793>.
- Xu, X., Y. Chen, and J. Miao. 2024. "Opportunities, Challenges, and Future Directions of Large Language Models, Including ChatGPT in Medical Education: A Systematic Scoping Review." *Journal of Educational Evaluation for Health Professions* 21:6. <https://doi.org/10.3352/jeehp.2024.21.6>.