Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 5 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i05.53529

PAPER

Exploring Medical Caption Generation through OpenAI's ChatGPT-4 Model: A PRISMA Review

Salma Elgayar[1](✉), Ibrahim I. M. Manhrawy[2], Amro M. Soliman[3], Safwat Hamad[4], El-Sayed M. Horbaty[5]

[1]Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

[2]Faculty of Information Technology, Applied Science Private University, Amman, Jordan

[3]Special Education Department, College of Education, King Khalid University, Abha, Saudi Arabia

[4]Business Analytics Department, School of Economic and Business Administration (SEBA), Saint Mary's College of California, Moraga, CA, USA

[5]Scientific Computing, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

salma.elgayar@cis.asu.edu.eg

ABSTRACT

This study explores the importance of the ChatGPT-4 model in medical caption generation, its advantages, applications, and limitations, using a PRISMA strategy on Medline and PubMed medical datasets to extract relevant studies from over a year ago concerning "ChatGPT" and "Medical Report Generation." The search employed keywords such as ("ChatGPT" OR "GPT model") AND ("medical caption generation" OR "medical image captioning" OR "radiology captioning"). The PRISMA search strategy led to the selection of seven promising papers. We conducted a brief comparison among the selected papers, taking into account their key focus, the datasets used, the models evaluated, the research results, and the challenges highlighted. Additionally, ChatGPT4's performance was evaluated by uploading sample medical images from different dataset modalities such as PathVQA, VQA-Med 2020, RadioGraphy Captions (RGC), and Radiology Objects in Context (ROCO) to establish whether it could generate coherent and contextually correct medical captions as true outputs and correctly answer medical questions with output performance BLEU = 0.5012 and ROUGE-L = 0.8000 scores. This study provides state-of-the-art evidence that ChatGPT demonstrates remarkable performance in report generation and answering medical questions under supervision.

KEYWORDS

medical field, ChatGPT, artificial intelligence, PRISMA, medical image caption generation

1 INTRODUCTION

OpenAI's GPT-4 model has made significant progress in many fields, with healthcare being one of the most important. Symbolic logic, a key part of classical artificial intelligence, works by following step-by-step logical rules. Before it can learn, certain important traits need to be identified. Experts are often needed to create databases and define logical rules for these systems, which act as expert advisors in various fields. However, symbolic logic has limitations, such as requiring regular maintenance and struggling to adapt to changing environments.

Elgayar, S., Manhrawy, I.I.M., Soliman, A.M., Hamad, S., Horbaty, El.-S.M. (2025). Exploring Medical Caption Generation through OpenAI's ChatGPT-4 Model: A PRISMA Review. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(5), pp. 18–30. https://doi.org/10.3991/ijoe.v21i05.53529

Article submitted 2024-11-26. Revision uploaded 2025-02-05. Final acceptance 2025-02-05.

On the other hand, the perceptron paradigm, which inspired the development of neural networks, is based on how neurons communicate in living organisms. Neural networks can combine information from many different sources, making them more flexible and scalable. The procedure of practical development involves an extensive amount of learning data, and it is suitable for challenging decision-making models, such as deep learning. Deep learning models solve complex problems by identifying features through training on large databases and using multilayer neural network architectures, achieving remarkably successful results.

A generative language model is called a generative pre-trained transformer (GPT) or ChatGPT [1], which is a category of large language models (LLMs) [2] derived from deep learning and the generative artificial intelligence framework. Using artificial intelligence trains itself on the structure and semantics of language while using multi-layer attention processes to process new data during training. The primary intent of ChatGPT is to understand and generate natural languages. This is achieved by training in a huge text corpora and large-scale parameterization. This had enabled ChatGPT to gain the ability of learning and generating text and hence making achievable a really large number of applications, such as text classification, information summary, even in task-based learning applications [3], and translations into other languages [4]. The potential of ChatGPT-4 has attracted a lot of attention from the healthcare industry [5], particularly for the application in smart medical diagnosis [6]. This includes the application in the clinical environments for managing medical records, surgical reports, radiological results [7], and summary discharges [8]. Traditional healthcare record keeping in clinical practice requires organized patient progress tracking over time and comparing textual data scattered over several files. In the same way, organizing radiological results, surgical reports, and associated documents requires the efficient generation of templates and the elimination of unnecessary details. Case reports are also a demanding application of clinical data. Regardless of many possible advantages, questions have been raised about the accuracy of the output material by ChatGPT [9], the ethics of the academic world [10], privacy, security, and possible biases in training databases. These are very important key factors to take into consideration, given the sensitivity of medical information and the necessary accuracy in medical documentation [11]. Case reports have been an essential part of the medical literary works since they distribute information about unique medical experiences, treatment results, and newly discovered illnesses. Examples of real-life applications of ChatGPT-4 in clinical assistance include providing decision-making support to physicians by using input symptoms to suggest possible diagnoses or therapies. Writing medical records requires careful planning, clear writing, and a detailed, step-by-step account of a patient's illness. However, there is still an absence of thorough study on the precise techniques that medical professionals use ChatGPT for in generating case reports, as well as the scope to which ChatGPT affects both the accuracy and the quality of these reports.

1.1 Research Questions

The rapid advancements in AI, particularly with GPT-4, have opened new possibilities in medical image captioning. However, its diagnostic reliability and

potential for clinical application remain underexplored. To address this, we focus on the following research questions:

1. Can ChatGPT-4 generate diagnostically accurate captions for radiology images?
2. What are the key limitations of ChatGPT-4 in medical image captioning, and how can they be addressed?

The main contributions of this work are listed as follows:

- An exhaustive review of applications, advantages, and limitations of the usage of ChatGPT-4 in the healthcare field.
- The PRISMA methodology to establish the most recent study on the role of ChatGPT-4 in medical image caption generation.
- Evaluation of ChatGPT-4 performance by uploading medical images for the goal of testing the output to medical questions and the quality of the generated captions.

2 CHATGPT-4 ASSISTANCE AND APPLICATIONS

ChatGPT-4 has many different applications in healthcare, providing intelligent support across various domains:

- **Medical Documentation [12]:** Helps automate clinical documentation, summarizing patient visits, and reduces administrative workload for physicians in the healthcare system.
- **Virtual Health Assistants:** Provides patients with personalized support, from answering health-related questions to reminding them of their appointments and tracking their symptoms.
- **Health Evaluation and Decision-making [13]:** Helps users in symptom analysis and gives an initial recommendation and further guidance on healthcare, improving early diagnosis.
- **E-health Support [14]:** Makes virtual consultations more efficient for physicians and patients by offering them real-time assistance during e-health visits.
- **Medical Research and Training [15]:** Aids in approaching medical literature, helps review research data, and provides interactive training for both students and professionals in medicine.
- **Mental Health Support [16]:** Shares conversations with patients in order to give them emotional support, coping strategies, and connections to mental health resources.
- **Medication Information [17]:** Provides accurate and convenient medicine information for both patients and professionals regarding medicine interactions, side effects, and dosage information.

Below are some of the advantages, applications, and limitations of OpenAI's GPT-4 model in the medical field summarized in Table 1.

Table 1. ChatGPT-4's advantages, applications, and limitations

Category	Key Points
Advantage	• Accessibility • Rapid Information Retrieval • Personalized Health • Education • Scalability • Multilingual Support • Cost Efficiency
Application	• Medical Documentation • Virtual Health Assistants • Symptom Checking and Triage • Telemedicine Support • Medical Research and Training • Mental Health Support • Drug Information
Limitation	• Not a Replacement for Human Expertise • Ethical and Privacy Concerns • Require further Fine-Tuning

Another important employment of GPT-4 in the medical field is the development of virtual assistance to assist patients with self-management of their health. For physicians and nurses, it may be used to create automated summaries of patient meetings and medical histories, which would make maintaining medical records easier.

Medical staff can utilize ChatGPT-4 to take their medical remarks and extract relevant information from patient data, including imaging reports or lab results, and even automatically outline the main points such as symptoms, diagnosis, and treatment. It can also assist in the recruitment of clinical trials by reviewing a large volume of patient information to identify individuals meeting the requirements of a trial's eligibility criteria.

3 RESEARCH METHODOLOGY

This study is based on search, screening, exclusion, and inclusion protocols based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) [18] guidelines as presented in Figure 1. A literature search of medical databases Medline and PubMed was made in November 2024 for articles published in the last year that described the use of ChatGPT in generating medical reports, focusing on "ChatGPT" and "Medical Report Generation," as these databases cover a large number of biomedical and health-related research. The medical database search used these keywords: ("ChatGPT" OR "GPT model") AND ("medical caption generation" OR "medical image captioning" OR "radiology captioning").

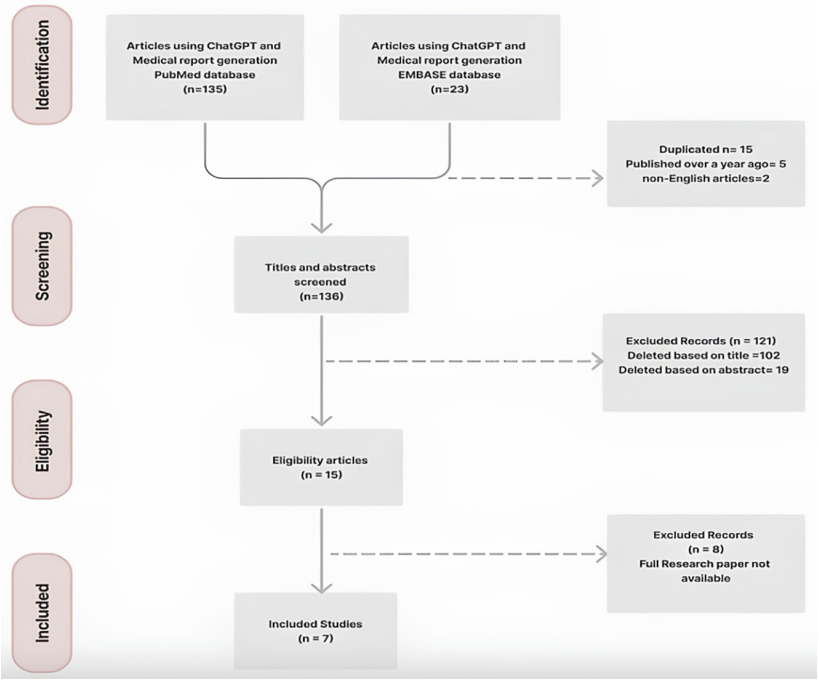

Fig. 1. PRISMA strategy for research using ChatGPT in medical report generation

3.1 Research Selection

An objective review of the abstracts and titles, as presented, was completed by the authors to assess the relevance of each study. Full articles of the selected studies were then reviewed thoroughly for their carrying out in the inclusion criteria.

The criteria had been used to ensure that only research applicable to the ChatGPT-based objectives was considered. In addition, exclusion criteria were considered to filter out papers that didn't fit the review's parameters in Table 2.

Table 2. Selection criteria

Inclusion Criteria
• Paper published in English language. • Paper published in a year. • Papers used ChatGPT in medical report generation. • Research paper matched with search keywords ("ChatGPT" OR "GPT model") AND ("medical caption generation" OR "medical image captioning" OR "radiology captioning").
Exclusion Criteria
• Full article of the study is not available, only the abstract. • Study is in a non-English language. • Study uses ChatGPT in the medical field but not for report generation. • Study is older than one year.

3.2 Results

To explore the utilization and limits of ChatGPT in medical caption generation, we classified key aspects of the selected studies. The seven studies were systematically categorized based on their **focus area, application domain, methodology, and type of model employed**, summed up in Table 3.

Table 3. Comparison of the key focus, datasets, models, outcomes, and challenges of each study

Paper Title	Key Focus	Dataset Used	Models Evaluated	Main Outcomes	Challenges Highlighted
MED-ChatGPT CoPilot [19]	Medical Text Data Mining with ChatGPT4	306 Medical papers	GPT-4 preview, ChatGPT with Prompt Engineering	Enhanced ChatGPT performance by 7.90%	Require a larger medical knowledge base
ChatGPT in healthcare: A taxonomy and systematic review [20]	ChatGPT in Healthcare Review	PubMed Database	ChatGPT	ChatGPT achieves moderate performance	ChatGPT not purposive for clinical deployment
Large language models in medical and healthcare fields: application advances and challenges [21]	LLMs in HealthcareApplications	175 studies	Various LLMs	Comprehensive review of LLMs in healthcare	Challenges comprise data security, bias, fairness, accuracy
Medical image captioning via generative pretrained transformes [22]	Clinical Image Caption Generation	Open-I, MIMIC-CXR, MS-COCO	Show-Attend-Tell, GPT-3	Efficient for chest X-ray image captioning	N/A
Multimodal ChatGPT for Medical Applicatios: Experimental Study of GPT-4V [23]	GPT-4V in Medical Visual Question Answering (VQA)	Pathology, Radiology Datasets (11 modalities)	GPT-4V	GPT-4V found unreliable for real-world diagnostics	Accuracy performance is 50%
Practical Evaluation of ChatGPT Performance for Radiology Report [24] Generation	ChatGPT for Radiology Report Generation	MIMIC Chest X-ray Database	ChatGPT, Bart, XLM, DeBERTa	Bart and XLM perform highly, mirroring physician reports	Major change in NLP model performance
The Utility of ChatGPT in Diabetic Retinopathy Risk Assessment: A Comparative Study with Clinical Diagnosis [25]	ChatGPT in Clinical Practice	1-Diabetic Retinopathy Clinical and Biochemical Dataset. 2-Dataset of 111 patients with diabetes.	ChatGPT4	Reliability0.91 in predicting diabetic risk but average 67–73% sensitivity & 54–68% specificity and fair agreement with clinical diagnoses 0.351 Cohen's kappa	1-Sensitivity and Specificity. 2-Optimization Necessarily.

Table 4. Comparative analysis of 7 research papers under different categories

Paper Category/Paper Reference	[19]	[20]	[21]	[22]	[23]	[24]	[25]
Focus Area Distribution							
Medical Text Mining	x						
Medical Decision Support							x
Radiology Report Generation						x	
Clinical Image Caption Generation/QA				x	x		
Systematic Review		x	x				
Application Domain Distribution							
General Health Care/NLP	x	x	x				x
Medical Imaging & Radiology				x	x	x	
Methodology Distribution							
Experimental Studies	x			x	x	x	x
Survey/Review		x	x				
Type of Model Used							
GPT Based Model	x	x		x	x	x	x
Other LLMs			x				

Table 4 clearly compares seven research papers classified based upon focus area, application domain, methodology, and model type. Some of the noticeable observations are as follows:

- It is clear from that the most spotted category in the literature is GPT-based models, with five papers.
- The type of survey/review, medical imaging and radiology, and general healthcare/NLP also have lots of representations, with four papers each.
- Number of papers deal with experimental studies and clinical image caption generation/QA three papers each.
- Less shown, with only one or two papers each, are categories such as radiology report generation, medical decision support, and medical text mining.

Therefore, there is a robust focus on GPT-based models and their applications in the healthcare domain, but fewer studies are dedicated to specific areas of decision support or text mining.

4 EVALUATION OF CHATGPT'S PERFORMANCE IN MEDICAL IMAGE INTERPRETATION

We utilized several medical datasets, including the PathVQA Dataset [26], VQA-Med 2020 Training Set [27], RGC [28], and ROCO dataset [29], to evaluate the performance of ChatGPT-4V on understanding medical images. Picking one image from each dataset, we provide prompts to ChatGPT with either a question or generating a medical caption, then evaluating and testing how accurate and relevant its responses are with respect to the specific image.

- **PathVQA Dataset:** A dataset used to adapt the visual question answering task with an area in pathology. It consists of images of pathologic slides along with question/answer pairs to evaluate the extent to which a model can understand complex medical images and generate accurate and applicable medical answers.
- **VQA-Med 2020 Training Set:** A dataset is developed for testing VQA models on many of medical images in different modalities, such as X-rays and MRIs. Each question/answer has a medical image attached, hence forcing the model to understand the content in a medical context and retrieve an accurate answer.
- **RGC (Radiology Gross Anatomy Corpus):** This dataset contains labeled medical images, often used in tasks related to the detection of body structures. Useful for the evaluation of a model's ability to discover specific structures in images of radiology and to answer questions about regional anatomy correctly.
- **ROCO Dataset:** ROCO is a dataset that has been very carefully prepared and is meant for the problems of visual question answering and image classification within the medical domain, with a focus on radiology images. It contains a diverse set of radiology images, such as X-rays, MRIs, and CT scans to annotate with detailed descriptions and contextual information.

Therefore, it allows for evaluating models on various tasks requiring an understanding of specific content, terminology, and diagnostic reasoning in radiology.

The performance of ChatGPT-4 in working on medical images from various datasets is depicted in Figures 2 to 5: PathVQA, VQA-Med 2020, RGC, and ROCO.

It compares true captions with highly specific medical terminology and details to the AI predictions for each dataset. While ChatGPT-4 shows a reasonable understanding of medical imaging, precision and completeness don't align with the truth captions. It thus has the potential to assist in medical image interpretation. However, it needs optimization to achieve clinical-grade accuracy and reliability.

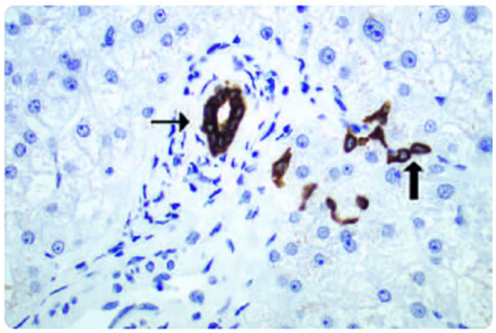

Fig. 2. A sample image from the PathVQA dataset

Answer: In the canals of hering.

Fig. 3. A sample image from the VQA-Med 2020 dataset

Answer: Chance-type fracture.

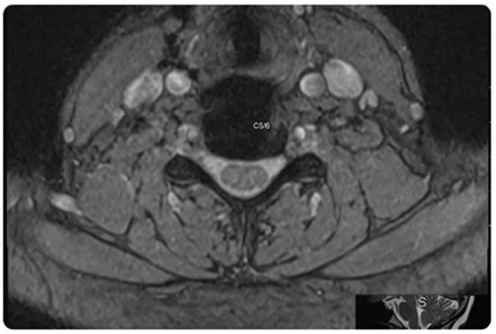

Fig. 4. A sample image from RGC dataset

True caption: MRI, no disc space narrowing, disc protrusion, facet arthropathy, canal stenosis, or foraminal narrowing.

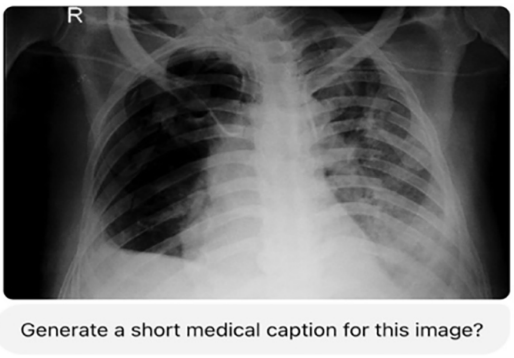

Fig. 5. A sample image from ROCO dataset

True caption: Chest X-ray anteroposterior right subclavian central venous catheter along the right border of mediastinum and left subclavian.

4.1 Results and Analysis of GPT-4 Generated Captions

We have calculated the BLEU and ROUGE-L scores over the previously generated medical captions in Figures 2–5 to assess how well medical image captions developed by GPT-4 are accurate. While ROUGE-L measures the longest common subsequence, reflecting mostly structural similarity, BLEU calculates n-gram overlap between generated and reference captions.

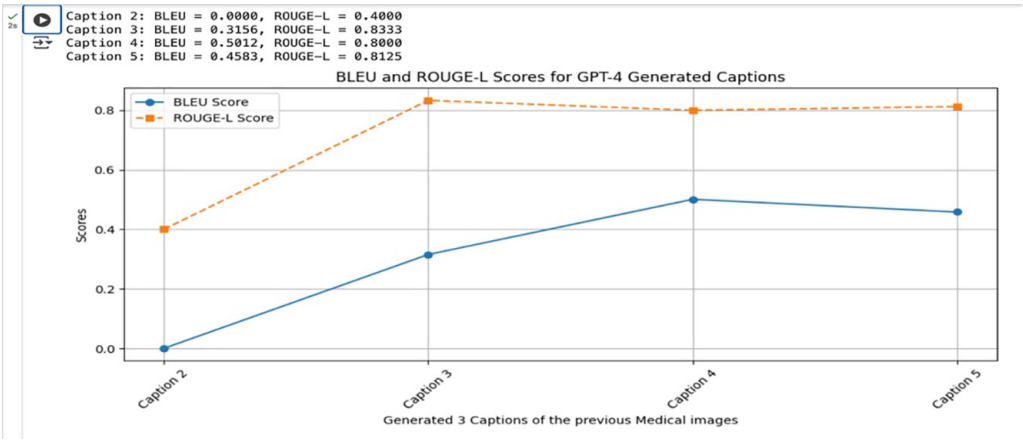

Fig. 6. BLEU in solid blue and ROUGE-L in orange for 4 generated captions

The BLEU in Figure 6, therefore, starts at Caption 2 and then increases to 0.3156 for Caption 3 and reaches a peak of 0.5012 for Caption 4 before slightly decreasing to 0.4583 for Caption 5. This trend indicates that the lexical similarity between the generated and reference captions improves initially but fluctuates slightly in later captions, possibly due to variations in GPT-4's captioning approach.

The trend is different for ROUGE-L scores: from 0.4000 for Caption 2 to a high 0.8333 for Caption 3 and stabilization around 0.8000 for Captions 4 and 5.

This suggests GPT-4 keeps strong structural coherence with the reference captions, though the exact lexical matches—as measured by BLEU—have weakened.

These differences between BLEU and ROUGE-L scores show that, while GPT-4-generated captions mostly maintained overall meaning and structure from the reference captions, they often used different phrasing. The BLEU declines on Caption 6, with its high ROUGE-L score, would indicate that GPT-4 generates paraphrases that are different yet correctly convey the meaning without exact n-gram matches.

5 DISCUSSION AND LIMITATIONS

The model shows a strong performance with accurate answers for the question/answering objectives of both PathVQA and VQA-Med 2020. This indicates that ChatGPT can process and respond with a fairly high degree of accuracy to specific medical question-answering tasks. Furthermore, ChatGPT will mainly correctly provide descriptive information about the anatomical structures in medical images when attempting to generate captions for the RGC and ROCO medical datasets.

However, the generated captions show limited diagnostic specificity, with BLEU scores ranging to 0.50 and ROUGE-L scores between 0.40 and 0.83, indicating a need for more detailed clinical interpretation, relating only to the general evaluation content of these medical images. The recognized difference between the model's performance on PathVQA, VQA-Med 2020, and RGC, ROCO could simply result from the availability of the datasets. PathVQA and VQA-Med 2020 are publicly available on Kaggle, and it can provide ChatGPT with more accurate information. The RGC and ROCO are private datasets; therefore, the chances of ChatGPT generating a detailed diagnosis for an image based on a source other than human medical supervision or additional training on the datasets are even more limited. This highlights the bias in data availability; ChatGPT is capable of performing detailed medical image interpretation. The only way that ChatGPT-4 is going to become a more effective tool in the field of radiology and pathology is through integration with the medical community. At that time, the fine-tuning of its responses will be possible, capturing detailed features and diagnostic accuracy with the presence of medical experts.

First, using a larger data set with more diverse characteristics combined with better prompting techniques and expert feedback would likely enhance diagnostic performance with GPT-4V.

Second, working in a collaborative approach with input from both radiologists and pathologists could help improve the current limitations and guide the model to a further improvement.

Most of the information in this dataset comes from image-question pairs and lacks detailed clinical history or images from different angles. Adding this kind of detailed information would better reflect the complete approach used by medical professionals and likely improve the model's accuracy and relevance.

6 CONCLUSION

The use of the ChatGPT application and other large language models in the domain of medical caption generation is an extremely promising area across various healthcare fields. These modes increase diagnosis accuracy, reduce the administrative load, smooth the workflow, and provide structured, insightful captions for medical images.

Across the reviewed studies, significant advancements have been highlighted, including ChatGPT's capability to integrate medical knowledge bases for aiding in

clinical decision-making (MED-ChatGPT CoPilot) and its potential to improve radiology reporting with clearer and diagnostically accurate captions (AI-driven Medical Captioning for Radiology).

The integration of multimodal data, as explored in Multimodal Integration of ChatGPT for Medical Image Interpretation, has shown how natural language processing can blend with image analysis to enrich medical reporting.

Despite these developments, several challenges remain. The risks of overreliance on AI-generated content and the need for human oversight continue to be expressed (ChatGPT and Medical Documentation).

Additionally, the full scalability and real-world clinical testing of these models, particularly across the diverse healthcare settings, remain largely unachieved (A Systematic Review of ChatGPT in Healthcare).

Although these technologies hold great potential to transform medical caption generation, additional research is required to overcome their limitations, especially on the accuracy, integration with the clinical systems, and ethical concerns related to AI in healthcare (Medical Large Language Models in Practice).

Overall, ChatGPT and similar models in its class are expected to play a significant role in the future of medical documentation and reporting, but they need to be implemented very cautiously with ongoing refinement.

7 ACKNOWLEDGMENT

Acknowledgments: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University. The funding is provided by the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups [grant number RGP.2/426/45].

8 REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[2] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] I. Lytovchenko, Y. Lavrysh, O. Synekop, V. Lukianenko, O. Chugai, and I. Shastko, "The use of ChatGPT in task-based ESP learning at university: Does it make a difference?" *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 2, pp. 4–22, 2025. https://doi.org/10.3991/ijim.v19i02.51115

[4] H. Hassani and E. S. Silva, "The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field," *Big Data and Cognitive Computing*, vol. 7, no. 2, p. 62, 2023. https://doi.org/10.3390/bdcc7020062

[5] E. Loh, "ChatGPT and generative AI chatbots: Challenges and opportunities for science, medicine and medical leaders," *BMJ Leader*, vol. 8, no. 1, pp. 51–54, 2024. https://doi.org/10.1136/leader-2023-000797

[6] H. Mehta, V. Shah, S. Mishra, N. Swain, C. Kulkarni, and D. Swain, "Smart diagnosis: Leveraging machine learning for early detection of hepatitis in healthcare," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 1, pp. 26–40, 2025. https://doi.org/10.3991/ijoe.v21i01.51383

[7] A. Rau *et al.*, "A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR appropriateness guidelines," *Radiology*, vol. 308, no. 1, 2023. https://doi.org/10.1148/radiol.230970

[8] S. Singh, A. Djalilian, and M. J. Ali, "ChatGPT and ophthalmology: Exploring its potential with discharge summaries and operative notes," *Semin. Ophthalmol.*, vol. 38, no. 5, pp. 503–507, 2023. https://doi.org/10.1080/08820538.2023.2209166

[9] S. Ariyaratne, P. Karthikeyan Iyengar, N. Nischal, N. Chitti Babu, and R. Botchu, "A comparison of ChatGPT-generated articles with human-written articles," *Skeletal Radiol.*, vol. 52, no. 9, pp. 1755–1758, 2023. https://doi.org/10.1007/s00256-023-04340-5

[10] E. Fournier-Tombs and J. McHardy, "A medical ethics framework for conversational artificial intelligence," *J. Med. Internet Res.*, vol. 25, p. e43068, 2023. https://doi.org/10.2196/43068

[11] K. Parmar and I. K. Verma, "A comprehensive approach to enhancing doctor-patient interaction: Bridging the gap for better healthcare," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 13, pp. 4–23, 2024. https://doi.org/10.3991/ijoe.v20i13.50345

[12] J. Liu, C. Wang, and S. Liu, "Utility of ChatGPT in clinical practice," *J. Med. Internet Res.*, vol. 25, p. e48568, 2023. https://doi.org/10.2196/48568

[13] H. Fraser, D. Crossland, I. Bacher, M. Ranney, T. Madsen, and R. Hilliard, "Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study," *JMIR Mhealth Uhealth*, vol. 11, p. e49995, 2023. https://doi.org/10.2196/49995

[14] T. M. Alanzi, "Impact of ChatGPT on teleconsultants in healthcare: Perceptions of healthcare experts in Saudi Arabia," *J. Multidiscip. Healthc.*, vol. 16, pp. 2309–2321, 2023. https://doi.org/10.2147/JMDH.S419847

[15] Y. Wu, Y. Zheng, B. Feng, Y. Yang, K. Kang, and A. Zhao, "Embracing ChatGPT for medical education: Exploring its impact on doctors and medical students," *JMIR Med. Educ.*, vol. 10, p. e52483, 2024. https://doi.org/10.2196/52483

[16] P. Raile, "The usefulness of ChatGPT for psychotherapists and patients," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, p. 47, 2024. https://doi.org/10.1057/s41599-023-02567-0

[17] B. Morath *et al.*, "Performance and risks of ChatGPT used in drug information: An exploratory real-world analysis," *European Journal of Hospital Pharmacy*, vol. 31, no. 6, pp. 491–497, 2024. https://doi.org/10.1136/ejhpharm-2023-003750

[18] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med*, vol. 6, no. 7, p. e1000097, 2009. https://doi.org/10.1371/journal.pmed.1000097

[19] W. Liu, H. Kan, Y. Jiang, Y. Geng, Y. Nie, and M. Yang, "MED-ChatGPT CoPilot: A ChatGPT medical assistant for case mining and adjunctive therapy," *Front. Med. (Lausanne)*, vol. 11, p. 1460553, 2024. https://doi.org/10.3389/fmed.2024.1460553

[20] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, "ChatGPT in healthcare: A taxonomy and systematic review," *Comput. Methods Programs Biomed.*, vol. 245, p. 108013, 2024. https://doi.org/10.1016/j.cmpb.2024.108013

[21] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: Applications, advances, and challenges," *Artif. Intell. Rev.*, vol. 57, no. 11, p. 299, 2024. https://doi.org/10.1007/s10462-024-10921-0

[22] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, "Medical image captioning via generative pretrained transformers," *Sci. Rep.*, vol. 13, no. 1, p. 4171, 2023. https://doi.org/10.1038/s41598-023-31223-5

[23] Z. Yan, K. Zhang, R. Zhou, L. He, X. Li, and L. Sun, "Multimodal ChatGPT for medical applications: An experimental study of GPT-4V," *arXiv preprint arXiv:2310.19061*, 2023.

[24] M. Soleimani, N. Seyyedi, S. M. Ayyoubzadeh, S. R. N. Kalhori, and H. Keshavarz, "Practical evaluation of ChatGPT performance for radiology report generation," *Acad. Radiol.*, vol. 31, no. 12, pp. 4823–4832, 2024. https://doi.org/10.1016/j.acra.2024.07.020

[25] E. Fikri, "The utility of ChatGPT in diabetic retinopathy risk assessment: A comparative study with clinical diagnosis [Letter]," *Clin. Ophthalmol.*, vol. 18, pp. 127–128, 2024. https://doi.org/10.2147/OPTH.S457160

[26] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "Pathological visual question answering," *TechRxiv*, 2020. https://doi.org/10.36227/techrxiv.13127537.v1

[27] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, and H. Müller, "Overview of the VQA-med task at imageCLEF 2021: Visual question answering and generation in the medical domain," 2021.

[28] L. Xu, B. Liu, A. H. Khan, L. Fan, and X.-M. Wu, "Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark," in *Conference on Health, Inference, and Learning*, PMLR, 2023, pp. 117–132.

[29] J. Rückert *et al.*, "ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset," *Sci. Data*, vol. 11, p. 688, 2024. https://doi.org/10.1038/s41597-024-03496-6

9 AUTHORS

Salma Elgayar is a Lecturer Assistant at the Faculty of Computer and Information Science, Ain Shams University. She has extensive research experience in Artificial Intelligence (AI), Deep Learning, and Machine Learning (ML) technologies. Her work focuses on advancing intelligent systems and their applications in various domains (E-mail: salma.elgayar@cis.asu.edu.eg).

Ibrahim I. M. Manhrawy received the B.Sc. (Hons.) in Mathematics and Computer Science and the M.Sc. in Computer Science from Menoufia University, Egypt, in 2006 and 2015, respectively, and Ph.D. in machine learning at the Faculty of Science, Menoufia University, Egypt, in 2021, He is currently employed School of Mathematics and Computer Science. He has many publications in the fields of machine learning. His research interests include machine learning and optimization (E-mail: i_manharawy@asu.edu.jo).

Amro M. Soliman currently working as an Associate Professor of Mental Health at King Khalid University, Abha - Saudi Arabia, received his Ph.D. from Ain-Shams University, Cairo, Egypt, in 2013, M.Sc. from Ain-Shams University, Cairo, Egypt, in 2009, and B.Sc. from Ain-Shams University, Cairo, Egypt, in 2000. His research interests include sign language for the deaf, mental disability, mental health, and Using technologies for people with special needs (E-mail: Amro@kku.edu.sa).

Safwat Hamad is a Professor of AI and Data Science, ITIL Expert, and IT Principal Consultant at Saint Mary's College of California and Ain Shams University. He is an experienced research and outreach specialist with contemporary insight into Quantum Computing, Artificial Intelligence (AI), Machine Learning (ML), Robotic Process Automation (RPA), Interconnected Intelligence, High-Performance Computing (HPC), Advanced Analytics, Cybersecurity, and associated Business and Digital Transformation technologies (E-mail: Shh4@stmarys-ca.edu).

Prof. El-Sayed M. Horbaty is a Professor of Scientific Computing at the Faculty of Computer and Information Sciences, Ain Shams University. His current areas of research include distributed and parallel computing, cloud computing, image processing, e-health computing, and optimization of computing algorithms. His work has been published in numerous journals, including *Parallel Computing*, *International Journal of High Performance and Grid Computing*, *International Journal of Information Security*, *International Journal of Computer Communication and Information Security*, and *International Review on Computers and Software* (E-mail: shorbaty@cis.asu.ed.eg).