Letter to the Editor

ChatGPT in radiology structured reporting: analysis of ChatGPT-3.5 Turbo and GPT-4 in reducing word count and recalling findings

Carlo A. Mallio[1,2]^, **Caterina Bernetti**[1,2], **Andrea C. Sertorio**[1,2], **Bruno Beomonte Zobel**[1,2]

[1]Fondazione Policlinico Universitario Campus Bio-Medico, Rome, Italy; [2]Research Unit of Radiology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy

Correspondence to: Carlo A. Mallio, MD, PhD. Fondazione Policlinico Universitario Campus Bio-Medico, via Álvaro del Portillo, 200, 00128 Rome, Italy; Research Unit of Radiology, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy. Email: c.mallio@policlinicocampus.it.

Submitted Sep 11, 2023. Accepted for publication Nov 20, 2023. Published online Jan 05, 2024.
doi: 10.21037/qims-23-1300
View this article at: https://dx.doi.org/10.21037/qims-23-1300

Introduction

Artificial intelligence (AI) is a field of computer science based on systems and algorithms capable of performing tasks that typically require human intelligence, such as learning, problem-solving, and decision-making (1,2).

ChatGPT is a natural language processing (NLP) tool designed for understanding and generating human-like text. In fact, it is a remarkable example of large language models (LLMs) which has garnered attention due to its conversational abilities and competence in NLP, proving to be capable of generating coherent and contextually relevant text responses (3-5). Indeed, ChatGPT has recently emerged as a revolutionary tool with potential applications in various domains, including healthcare and radiology. In this context, total-body computed tomography (CT) is one of the most exploited and versatile diagnostic examinations that allows to evaluate a patient's brain and body in extreme detail. However, CTs produce a considerable amount of data that must be accurately interpreted and communicated (6).

Radiologists are responsible for image analysis and reporting, which is often performed as unstructured free-text report. This kind of report is thorough and comprehensive, but also time-consuming and often burdened by several issues, including absence of standardized terminologies across institutions, difficult communication between healthcare professionals, recollection, categorization, and analysis of data for research and health management purposes (6).

Consequently, in the last years, there has been a growing interest in structured radiological reports to convert free-text information into systematized and standardized formats, making abundant information more accessible and organized, to improve radiological workflow, and possibly implement data extraction and analysis (7-9).

In the realm of radiology practice, AI and in particular LLMs, such as ChatGPT 3.5 Turbo and GPT-4, have shown extraordinary proficiency in comprehension, production, and manipulation of human language, demonstrating great potential also for the generation of structuring radiological reports (5,10-12). They are able to analyze lengthy and descriptive radiology reports and distill essential information into concise structured format. This function could improve overall efficiency, helping streamline workflows, implementing data extraction and providing explanations of radiological findings (13,14). However, it should be emphasized that formal validation of these applications and expert radiologist over-reading of structured reports generated by ChatGPT are mandatory,

^ ORCID: 0000-0002-0149-0801.

Table 1 Details on prompts provided to the LLMs, ChatGPT 3.5 Turbo and GPT-4

Group	Instruction (prompt)
1	Convert the following report of a total-body CT scan into a structured report, include as many details as possible. The format should be a table
2	Transform this free-text radiological report into a structured report. Includes only the essential information, so the structured report should be concise and organized in a table format
3	Transform this free-text radiological report into a structured report. The structured report should be concise, organized in table format, and contain only pathological elements

LLMs, large language models; GPT, generative pretrained transformer; CT, computed tomography.

due to possible errors (e.g., hallucinations or factual errors).

Indeed, these models can be applied to structured reporting in radiology but to what extent some findings might be overlooked by ChatGPT, while transforming reports, is currently not fully understood.

The study aims to provide valuable insights into the feasibility of LLMs, investigating the effectiveness of ChatGPT 3.5 Turbo and GPT-4 in transforming free-text radiological reports into structured formats, analyzing their ability of synthesizing, in terms of word count reduction, but also evaluating the quality of the structured report, based on recall rates of different categories of findings.

Methods

Ethical committee approval was not required due to the absence of patients and identifiable data involved. The data search was performed before March 29th, 2023.

We conducted this study using 60 fictitious total-body CT reports, randomly divided into three groups of 20 reports. The reports were created by consensus of two researchers: one radiologist (C.A.M., 12 years of experience) and one resident in radiology (A.C.S., 4 years of experience). All the reports were built as a close mirror of real-life total-body CT reports observed in clinical practice. All the analyses were performed in the Italian language and then translated in English only to provide example (Table S1).

Each group was subjected to different prompts proposed to ChatGPT 3.5 Turbo (OpenAI) and GPT-4 (OpenAI). We then collected and analyzed responses offered by the LLMs.

The aim of all prompts was to convert original free-text reports into a structured report, in a table format, with different levels of details and conciseness (*Table 1*):

❖ Prompt 1 (P1): urged the models to include as many details as possible in the structured report. The aim was to explore the level of detail and complexity of the content that the models could generate from the original report.

❖ Prompt 2 (P2): was more limited, requiring the models to select and include only essential information. The goal of this prompt was to determine how the model's filtered information and decide what is considered essential.

❖ Prompt 3 (P3): required the models to focus solely on the pathological elements present in the report. This served to understand the models' ability to correctly recognize and categorize medical anomalies.

The structured reports created by LLMs, were then analyzed, in comparison with the source free-text report, to verify the quality of information:

❖ A quantitative evaluation of the results, collecting and comparing word numbers in the original reports and in the structured reports generated by the models. The results, in terms of percentage of word count reduction, were considered as a direct measure of the effectiveness of the models in compacting information and reducing verbosity.

❖ A qualitative analysis of the structured reports extrapolated and proposed by the LLMs.

Due to the variability of classification systems in the literature, we decided to include a classification of the findings in the reports based on the CT Colonography Reporting and Data System (C-RADS) (15,16) (*Table 2*). The evaluation metric used to quantify effectiveness was defined as "recall" to estimate the ability of the models to keep or not different categories of findings from the free-text to the structured reports. The recall rate was calculated for each category of findings to understand various levels of competence in categorizing what is important and what is not. These metrics provide a comprehensive picture of the performance of the models, considering not only

Table 2 Explanation of the different C-RADS categories (15,16)

Category	Findings
C-RADS E1	Normal examination results or anatomical variants
C-RADS E2	Clinically unimportant findings
	No further evaluation or investigation indicated (e.g., renal cyst, diverticulosis)
C-RADS E3	Indeterminate, incompletely characterized, but probably benign
	Further clinical correlation and investigation could be carried out if indicated (e.g., minimally complex renal cyst)
C-RADS E4	Potentially important findings
	Further investigation required and communication to the referring physician, as per accepted practice guidelines (e.g., solid renal mass, abdominal aortic aneurysm)

C-RADS, CT Colonography Reporting and Data System; CT, computed tomography.

Table 3 Performance of GPT-3.5 Turbo in reducing word count while converting free-text into structured radiological reports

Group	Quantitative evaluation			Qualitative evaluation			
	Average no. of words in the original report	Average no. of words after transformation with GTP-3.5 Turbo	Reduction in words after transformation (%)	C-RADS E1 (%)	C-RADS E2 (%)	C-RADS E3 (%)	C-RADS E4 (%)
1	595.25	454.65	23.6	85	82	89	79
2	529.75	270.2	49.9	68	65	77	75
3	563.95	299.1	47	62	81	89	90

GPT, generative pretrained transformer; C-RADS, CT Colonography Reporting and Data System; CT, computed tomography.

their ability to correctly identify information, but also to exclude non-essential information and to balance these two competencies.

The qualitative analysis was performed by consensus of two researchers: one radiologist (C.A.M., 12 years of experience) and one resident in radiology (A.C.S., 4 years of experience).

Results

We analyzed for each LLMs, GPT-3.5 Turbo and GTP-4, the results of the three prompts proposed.

GPT-3.5 Turbo

For Group 1, which had the prompt (P1) to include as many details as possible, GPT-3.5 Turbo reduced the average word count by 23.6% compared to the original report, maintaining a good performance in terms of recall for the C-RADS categories, with a minimum of 79% for category E4 and a maximum of 89% for category E3.

For Group 2, which had the prompt (P2) to include only essential information, GPT-3.5 Turbo reduced the average word count by 49.9% compared to the original report, with a recall rate for the C-RADS categories ranging from 65% for category E2 to 77% for category E3.

For Group 3, which had the prompt (P3) to include only pathological elements, GPT-3.5 Turbo reduced the average word count by 47% compared to the original report, the recall for the C-RADS categories ranged from 62% for category E1 to 90% for category E4.

See *Table 3* for the complete results of GPT-3.5 Turbo.

GPT-4

For Group 1, GPT-4 reduced the average word count by 36.7% compared to the original report, with a recall for the C-RADS categories ranged from 80% for category E4 to 93% for category E2.

For Group 2, GPT-4 reduced the average word count by 75% compared to the original report, with a recall for the C-RADS categories ranging from 59% for category E1 to

77% for category E3.

For Group 3, GPT-4 reduced the average word count by 73.2% compared to the original report, with a recall for the C-RADS categories ranged from 36% for category E1 to 83% for category E4. See *Table 4* for the complete results of GPT-4.

Comment

The results of this study highlight the potential of LLMs such as ChatGPT 3.5 Turbo and GPT-4 in the process of structuring radiological reports. Both models demonstrated ability to transform free-text radiological reports into a structured format and reduce verbosity. Moreover, we observed that while transforming free-text into structured report some findings, even of potential clinical importance, might be missed. Furthermore, the results underscored differences between the two models, as well as between different report groups.

In Group 1, which received the prompt (P1) to include as many details as possible, both GPT-3.5 Turbo and GPT-4 showed good recall for all C-RADS categories, with GPT-4 reducing the word count more significantly compared to GPT-3.5 Turbo. This suggests that both models can generate detailed structured reports, but GPT-4 may be more effective in synthesizing information.

In Group 2, which received the prompt (P2) to include only essential information, GPT-4 reduced the word count by 75% compared to the original report but showed lower recall for the C-RADS E1 category compared to GPT-3.5 Turbo. This result suggests that while GPT-4 is effective in synthesizing reports, it may miss some information.

In Group 3, which received the prompt (P3) to include only pathological elements, both models showed good recall for the C-RADS E4 category, representing the most important findings. However, GPT-4 reduced the word count by 73.2% compared to the original report but showed lower recall for the C-RADS E1 category compared to GPT-3.5 Turbo. This indicates that GPT-4 can identify and effectively report pathological elements but may miss some important information.

In general, GPT-4 demonstrated a greater ability to reduce the number of words in reports compared to GPT-3.5 Turbo. This may be due to its more advanced architecture and its ability to generate more concise text. However, this greater synthesis capability often corresponded to a slightly lower recall rate in some categories of findings compared to GPT-3.5 Turbo, confirming that this synthesis could lead to overlooking findings.

Moreover, GPT-4 showed greater variation in recall scores across different C-RADS categories compared to GPT-3.5 Turbo. For example, in Group 3, GPT-4 showed a recall of 36% for category E1 and 83% for category E4. This result contrasts with GPT-3.5 Turbo, which showed a recall of 62% for category E1 and 90% for category E4. This could indicate that GPT-4 is more selective in reporting information, correctly focusing its attention on potentially pathological elements, and less on non-pathological ones, when required.

These differences highlighted that specific context and user needs can have an impact on the effectiveness of the models, and hence, must be considered when choosing the appropriate LLMs for the task required. For instance, where it is important to synthesize and focus on pathological elements, GPT-4 might be more suitable; whereas, in case of necessity to maintain as many details as possible, GPT-3.5 Turbo might be a better option.

Hence, it is fundamental to provide clear, specific, and well-crafted prompts to the models. This aligns with existing literature, which has highlighted the importance of prompting in guiding the behavior of language models (4,17-19).

The intentional variation of prompts performed in our study allowed us to evaluate the performance, in terms of flexibility and adaptability, of the LLMs under different conditions and to explore their ability to discern and categorize information based on generic terms. This is essential, as we are sought to understand how well these models, without specific training, can autonomously categorize medical information.

ChatGPT in the field of diagnostic imaging has been explored recently by a few papers. For instance, Adams *et al.* reported that GPT-4 can convert free-text into structured radiological reports with minor effort, possibly accounting for the challenges of structured reporting, facilitating standardization and data extraction (4).

Moreover, Lyu *et al.* (17) reported that ChatGPT can convert radiology reports into plain language with a good performance obtaining a score of 4.27 (five-point system) with 0.08 places of information missing and 0.07 places of misinformation.

While there is potential for these AI tools to aid in diagnosis, it's essential to understand the underlying mechanisms of their decision making. The algorithms of ChatGPT are trained on vast datasets and their decisions are based on patterns in the data rather than clinical understanding. Hence, while they can be incredibly

Table 4 Performance of GPT-4 in reducing word count while converting free-text into structured radiological reports

Group	Quantitative evaluation			Qualitative evaluation			
	Average no. of words in the original report	Average no. of words after transformation with GTP-3.5 Turbo	Reduction in words after transformation (%)	C-RADS E1 (%)	C-RADS E2 (%)	C-RADS E3 (%)	C-RADS E4 (%)
1	595.25	376.5	36.7	90	93	91	80
2	529.75	132.2	75	59	70	77	74
3	563.95	151.05	73.2	36	83	81	83

GPT, generative pretrained transformer; C-RADS, CT Colonography Reporting and Data System; CT, computed tomography.

accurate, there are some complicated clinical scenarios where they might be misled, especially if the training data is not representative. This nuanced understanding is crucial for any clinical application of AI.

Our findings, in fact, underline that radiologist's oversight is still needed while converting free-text into structured radiological reports in clinical practice. Moreover, we introduced the concept that missing findings should be classified according to the clinical importance, as we did in the present paper using C-RADS categories, to evaluate the performance of LLMs in the context of structured reporting.

It is important to mention that the use of LLMs in radiological practice is in the early stage and further research is needs since there are still several challenges to overcome. It would be interesting to explore how to improve the models' ability to recognize and report relevant information, how to adapt the models to the specific contexts and user needs, and how to effectively integrate the models in clinical practice and with existing health information management systems.

The results of our study, even though preliminary, suggest that these models can be useful tools for structuring radiology reports, having the potential to enhance the efficiency and effectiveness of radiological results communication, data analysis, and integration with health information management systems.

However, while ChatGPT offers many advantages in structuring radiology reports, it is crucial to understand its limitations, such as occasional oversights, for example missed findings of clinical relevance, or inaccuracies in reporting data. Since, the employment of these systems without proper validation and checks can introduce errors (e.g., hallucinations or factual errors) into the radiology reporting process, we must underline that human judgment is still crucial and nowadays LLMs must be considered as an aid rather than a replacement for human expertise.

Moreover, the integration of ChatGPT and similar AI tools in healthcare and radiology is not devoid of ethical considerations, regarding data privacy and the potential misuse of sensitive patient information. Hence, it is imperative to strike a balance between leveraging the capabilities of AI and ensuring the ethical treatment of patient data and care.

To this end, tools like ChatGPT must be utilized in a manner that ensures safety of patient-identifiable information. Regulatory standards, such as the Health Insurance Portability and Accountability Act (HIPAA), set stringent criteria for the protection of patient data (20). If ChatGPT is integrated into clinical workflows, institutions must ensure that these standards are, at least, met. Strategies to ensure patient data privacy when using ChatGPT can include hosting custom models locally, encrypting patient data, and using the tool in offline modes where feasible.

Lastly, ChatGPT holds great promise also as a tool for medical education and training toward the so-called self-directed learning, empowering students to enhance their skills and knowledge at their own pace (21). Once again, human supervision is still mandatory in this context, in order not to pass inaccurate or false information.

While the current applications of ChatGPT in radiology are encouraging, its future perspectives are intriguing. However, as we explore these future directions, risk of bias, data privacy protection, legal and ethical considerations must not be overlooked (9,10). Further research, including different fine-tuning strategies, is needed to explore how to optimize the use of LLMs and to understand if and eventually how they can be integrated into clinical practice.

Acknowledgments

Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1300/coif). C.A.M. serves as an unpaid editorial board member of *Quantitative Imaging in Medicine and Surgery*. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Greco F, Mallio CA. Artificial intelligence and abdominal adipose tissue analysis: a literature review. Quant Imaging Med Surg 2021;11:4461-74.
2. Sheng Y, Zhang J, Ge Y, Li X, Wang W, Stephens H, Yin FF, Wu Q, Wu QJ. Artificial intelligence applications in intensity modulated radiation treatment planning: an overview. Quant Imaging Med Surg 2021;11:4859-80.
3. Mallio CA, Bernetti C, Sertorio AC, Beomonte Zobel B. Large language models and structured reporting: never stop chasing critical thinking. Radiol Med 2023;128:1445-6.
4. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 2023;307:e230725.
5. Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. Radiol Med 2023;128:808-12.
6. Goel AK, DiLella D, Dotsikas G, Hilts M, Kwan D, Paxton L. Unlocking Radiology Reporting Data: an Implementation of Synoptic Radiology Reporting in Low-Dose CT Cancer Screening. J Digit Imaging 2019;32:1044-51.
7. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights Imaging 2020;11:10.
8. Granata V, Fusco R, Cozzi D, Danti G, Faggioni L, Buccicardi D, et al. Structured reporting of computed tomography in the polytrauma patient assessment: a Delphi consensus proposal. Radiol Med 2023;128:222-33.
9. Neri E, Granata V, Montemezzi S, Belli P, Bernardi D, Brancato B, et al. Structured reporting of x-ray mammography in the first diagnosis of breast cancer: a Delphi consensus proposal. Radiol Med 2022;127:471-83.
10. Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Radiology, structured reporting and large language models: who is running faster? Radiol Med 2023;128:1443-4.
11. Mallio CA, Napolitano A, Castiello G, Giordano FM, D'Alessio P, Iozzino M, Sun Y, Angeletti S, Russano M, Santini D, Tonini G, Zobel BB, Vincenzi B, Quattrocchi CC. Deep Learning Algorithm Trained with COVID-19 Pneumonia Also Identifies Immune Checkpoint Inhibitor Therapy-Related Pneumonitis. Cancers (Basel) 2021;13:652.
12. Mago J, Sharma M. The Potential Usefulness of ChatGPT in Oral and Maxillofacial Radiology. Cureus 2023;15:e42133.
13. Wang H, Cao J, Fan H, Huang J, Zhang H, Ling W. Compared with CT/MRI LI-RADS, whether CEUS LI-RADS is worth popularizing in diagnosis of hepatocellular carcinoma?-a direct head-to-head meta-analysis. Quant Imaging Med Surg 2023;13:4919-32.
14. Roca-Espiau M, Valero-Tena E, Ereño-Ealo MJ, Giraldo P. Structured bone marrow report as an assessment tool in patients with hematopoietic disorders. Quant Imaging Med Surg 2022;12:3717-24.
15. Lee SY, Landis MS, Ross IG, Goela A, Leung AE. Extraspinal findings at lumbar spine CT examinations: prevalence and clinical importance. Radiology 2012;263:502-9.
16. Zalis ME, Barish MA, Choi JR, Dachman AH, Fenlon HM, Ferrucci JT, Glick SN, Laghi A, Macari M, McFarland EG, Morrin MM, Pickhardt PJ, Soto J, Yee J; . CT colonography reporting and data system: a consensus proposal. Radiology 2005;236:3-9.
17. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, Wang G, Whitlow CT. Translating radiology reports into plain language using ChatGPT and GPT-4 with

prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6:9.

18. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 2023;104:269-74.
19. Buvat I, Weber W. Nuclear Medicine from a Novel Perspective: Buvat and Weber Talk with OpenAI's ChatGPT. J Nucl Med 2023;64:505-7.
20. Rosenbloom ST, Smith JRL, Bowen R, Burns J, Riplinger L, Payne TH. Updating HIPAA for the electronic medical record era. J Am Med Inform Assoc 2019;26:1115-9.
21. Ricotta DN, Richards JB, Atkins KM, Hayes MM, McOwen K, Soffler MI, Tibbles CD, Whelan AJ, Schwartzstein RM; . Self-Directed Learning in Medical Education: Training for a Lifetime of Discovery. Teach Learn Med 2022;34:530-40.

Cite this article as: Mallio CA, Bernetti C, Sertorio AC, Zobel BB. ChatGPT in radiology structured reporting: analysis of ChatGPT-3.5 Turbo and GPT-4 in reducing word count and recalling findings. Quant Imaging Med Surg 2024;14(2):2096-2102. doi: 10.21037/qims-23-1300