Enhancing Radiological Reporting in Head and Neck Cancer: Converting Free-Text CT Scan Reports to Structured Reports Using Large Language Models

Amit Gupta[1] Hema Malhotra[2] Amit K. Garg[3] Krithika Rangarajan[2]

[1] Department of Radiodiagnosis, All India Institute of Medical Sciences New Delhi, New Delhi, India
[2] Department of Radiology, Dr. Bhim Rao Ambedkar Institute Rotary Cancer Hospital, All India Institute of Medical Sciences New Delhi, India
[3] Indian Institute of Technology, New Delhi, India

Address for correspondence Krithika Rangarajan, MD, FRCR, Room No. 160D, Department of Radiology, Dr. Bhim Rao Ambedkar Institute Rotary Cancer Hospital, All India Institute of Medical Sciences New Delhi, Ansari Nagar 110029, New Delhi, India (e-mail: krithikarangarajan86@gmail.com).

Indian J Radiol Imaging 2025;35:43–49.

Abstract

Objective The aim of this study was to assess efficacy of large language models (LLMs) for converting free-text computed tomography (CT) scan reports of head and neck cancer (HNCa) patients into a structured format using a predefined template.

Materials and Methods A retrospective study was conducted using 150 CT reports of HNCa patients. A comprehensive structured reporting template for HNCa CT scans was developed, and the Generative Pre-trained Transformer 4 (GPT-4) was initially used to convert 50 CT reports into a structured format using this template. The generated structured reports were then evaluated by a radiologist for instances of missing or misinterpreted information and any erroneous additional details added by GPT-4. Following this assessment, the template was refined for improved accuracy. This revised template was then used for conversion of 100 other HNCa CT reports into structured format using GPT-4. These reports were then reevaluated in the same manner.

Results Initially, GPT-4 successfully converted all 50 free-text reports into structured reports. However, there were 10 places with missing information: tracheostomy tube ($n = 3$), noninclusion of involvement of sternocleidomastoid muscle ($n = 2$), extranodal tumor extension ($n = 3$), and contiguous involvement of the neck structures by nodal mass rather than the primary ($n = 2$). Few instances of nonsuspicious lung nodules were misinterpreted as metastases ($n = 2$). GPT-4 did not indicate any erroneous additional findings. Using the revised reporting template, GPT-4 converted all the 100 CT reports into a structured format with no repeated or additional mistakes.

Conclusion LLMs can be used for structuring free-text radiology reports using plain language prompts and a simple yet comprehensive reporting template.

Keywords
- GPT-4
- large language models
- structured reporting
- head and neck cancer

article published online
August 1, 2024

DOI https://doi.org/10.1055/s-0044-1788589.
ISSN 0971-3026.

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

Key Points

- Structured radiology reports in oncological patients, although advantageous, are not used widely in practice due to perceived drawbacks like interference with routine radiology workflow and scan interpretation.
- We found that GPT-4 is highly efficient in converting conventional CT reports of HNCa patients to structured reports using a predefined template.
- This application of LLMs in radiology can help in enhancing the acceptability and clinical utility of structured radiology reports in oncological imaging.

Summary Statement

Large language models can successfully and accurately convert conventional radiology reports for oncology scans into a structured format using a comprehensive predefined template and thus can enhance the utility and integration of these reports in routine clinical practice.

Introduction

Radiology reports serve a pivotal role in clinical decision-making, acting as the primary channel of communication regarding imaging findings. Traditional radiology reporting, often unstructured and dictated in a prose style, poses significant challenges for both human comprehension and computerized analysis. Structured reporting has been hailed as a solution to these challenges.[1] By standardizing the format and lexicon of reports, structured reporting not only enhances the clarity and completeness of the reports but also facilitates a more systematic review process during reporting.[2,3] With increasing patient access to their medical records, including radiology reports, the clarity and effectiveness of these reports become even more crucial. Structured reporting, by improving the communication of findings, can significantly enhance patient care.[4] Structured reports also allow for effective data mining, contributing significantly to research and the development of artificial intelligence (AI) based applications in radiology.[5] However, there are a few skepticisms against structured reporting including perceived restriction of radiologist's autonomy and nonstandard thinking, limited inclusion of details of the findings, interference with scan interpretation, cumbersome process to generate a structured report for complex cases, and, finally, behavioral resistance to divergence from the status quo.[1,6–8]

Recent advancements in natural language processing (NLP), particularly with deep learning algorithms and transformer-based models, have shown promising results in transforming unstructured information into structured data for various day-to-day applications like finance documents, inventory logs, customer preferences for businesses, and much more. In the clinical health care domain, various studies have demonstrated the practical applicability and effectiveness of these models.[9–14] The use of transformer models, such as the text-to-text transformer (T5) and domain-specific pretrained models, has been explored for generating structured reports from radiology free texts.[15–17] These models have shown remarkable performance in the extraction and standardization of relevant information, enabling the direct generation of machine-readable structured reports with minimal manual intervention.

In the context of head and neck cancers (HNCas), structured reporting can play an even more critical role. The radiology reports of these cancers requiring description of multiple sites, subsites, and individual structures in the neck are essential to guide effective treatment planning and follow-up.[18] However, structured reporting of these scans has not been widely adopted so far primarily due to lack of globally standardized imaging lexicon as well as their perceived limitations like interference with the radiologist's workflow.[3,8]

Thus, in this study, we aimed to extend the benefits of structured reporting to HNCa imaging by employing Generative Pre-Trained Transformer (GPT-4), a cutting-edge large language model (LLM) for converting free-text computed tomography (CT) scan reports into a predefined structured format, with minimal interference with the radiologist's workflow.

Materials and Methods

Study Design and Data Collection

This retrospective study was approved from ethical angle, and written informed consent was waived off by the Institutional Review Board. For inclusion in this study, we selected 150 CT reports of consecutive HNCa patients who had previously undergone CT at our institute between January and December 2021. These reports, initially in a free-text format, were divided according to five major subsites of primary tumor involvement including the oral cavity, oropharynx, hypopharynx, nasopharynx, and larynx. The retrieved CT reports were anonymized for further use in the study and all patient data (like unique hospital identification number, name, age, and sex) were removed from the text reports. These reports were then divided into two groups of 50 and 100 reports. The number of scans for each subsite included in the two groups of reports are mentioned in ►**Table 1**.

Table 1 Number of CT reports from different HNCa subsites included in the study

Site of primary tumor	Number of CT reports in group 1	Number of CT reports in group 2	Total
Oral cavity	4	7	11
Oropharynx	15	31	46
Hypopharynx	6	12	18
Nasopharynx	10	18	28
Larynx	15	32	47
Total	50	100	150

Abbreviations: CT, computed tomography; HNCa, head and neck cancer.

Development of Structured Reporting Template

A comprehensive structured reporting template for HNCa CT scans was developed by the authors (►**Supplementary Appendix 1**, available in the online version). This template was designed to encompass a detailed enumeration of various anatomical sites pertinent to HNCa, including specific subsites within these regions. Key imaging findings were systematically incorporated into the template, such as the status of cervical lymph nodes, any evidence of airway compromise, and the involvement of additional neck structures and vascular components. The template was crafted to ensure that all clinically relevant information for HNCa diagnosis, staging, and treatment planning was captured.

Phase 1: Initial Evaluation of GPT-4 for Structured Report Generation

Prompt Engineering

Multiple prompt variations were tested with few fictitious HNCa CT reports in the context window for GPT-4 to identify the most effective approach for structured report generation. Few prompts that were used as follows:

- Convert the free-text CT scan report into a structured report using the given template.
- Reformat the information given in the following free-text CT scan report into the provided structured template.
- Convert the following free-text CT scan report into a structured format strictly according to the provided template. Focus on detailing anatomical structures involved, including primary tumor location, size, and extent, as well as any involvement of lymph nodes, airways, and other neck structures within the template.

GPT-4 generated the structured reports with all these prompts. However, the responses with the first two prompts showed frequent deviation from the desired template with GPT-4 modifying the section and subsection names and descriptions. The last prompt demonstrated consistent results with the GPT-4 responses in following the desired template and was thus used further in this study.

Generation and Evaluation of Structured Reports

The selected best prompt was then used to convert the first group of 50 HNCa CT reports into a structured format using GPT-4 according to the developed template. The generated structured reports were subjected to a thorough evaluation and comparison with the original free-text reports by a radiologist (8 years of experience in body imaging). This objective assessment focused on three critical aspects: the presence of missing information, instances of misinterpreted information, and the identification of any additional erroneous details added by GPT-4 that were not part of the original free-text reports. This evaluation aimed to quantify the accuracy and completeness of the structured reports.

Phase 2: Evaluation of GPT-4 for Structured Report Generation using the Revised Template

Iterative Improvement and Template Refinement

Based on the insights gained from the initial evaluation, the structured reporting template was refined to explicitly address areas where inaccuracies or omissions were noted. The revised template with modifications marked in italics is shown in ►**Supplementary Appendix 2** (available in the online version).

Utilization of the Revised Template for Structured Report Generation

For testing the performance of GPT-4 with the final revised report template, the second group of 100 HNCa CT reports was used. The same prompt (as for the previous 50 reports) was used to convert these reports into a structured format using GPT-4 and the revised template. These generated reports were then objectively evaluated by the radiologist in the same manner as the previous structured reports. ►**Fig. 1** summarizes the study workflow.

Results

Performance of GPT-4 for Generating Structured Reports Using the Initial Template

Conversion of Free-Text Reports to Structured Format

In our study, GPT-4 was initially utilized to convert 50 free-text radiology reports of HNCa patients into a structured format. We observed that GPT-4 successfully transformed all provided reports, adhering to the predefined structured template specifically designed for HNCa imaging.

Identification and Analysis of Missing Information

Upon a detailed analysis of the converted reports, certain instances of missing information were identified, quantitatively summarized as follows:

- *Tracheostomy tube*: In three instances, the presence of a tracheostomy tube, a critical clinical detail, was omitted.
- *Anatomical detailing*: The structured reports lacked the description of the involvement of sternocleidomastoid muscle in two cases.

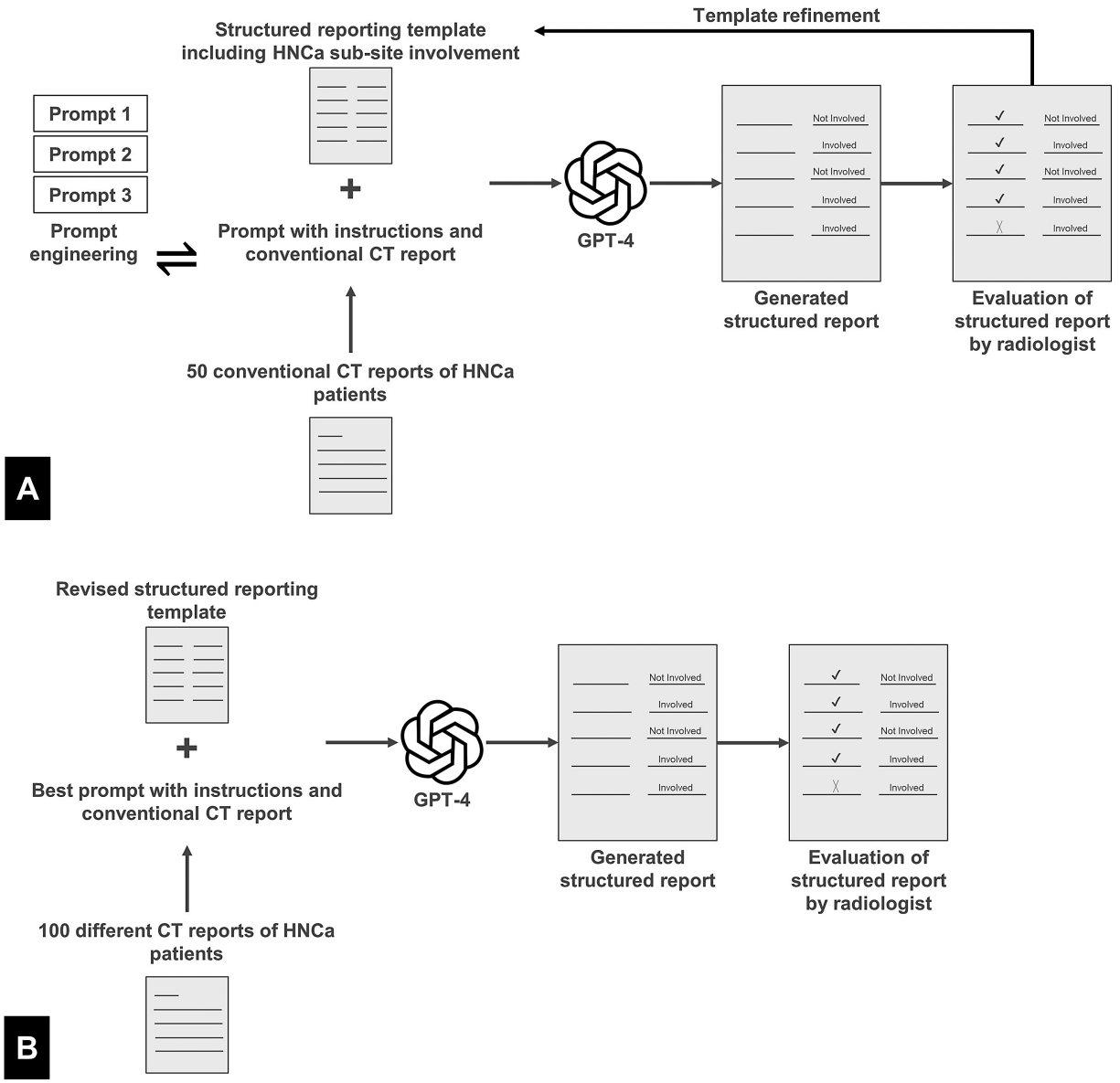

Fig. 1 Study workflow. (**A**) Initial evaluation of Generative Pre-trained Transformer 4 (GPT-4) for structured report generation using 50 head and neck cancer (HNCa) computed tomography (CT) reports followed by revision of the reporting template. (**B**) Testing of GPT-4 for conversion of 100 different HNCa CT reports into structured format using the revised reporting template.

- *Extranodal tumor extension*: Three reports failed to mention extranodal tumor extension, an essential factor in cancer staging and prognosis.
- *Contiguous involvement of neck structures*: In two reports, there was a missing distinction between the involvement of neck structures by nodal mass as opposed to primary tumor involvement.

Instances of Misinterpretation

In two instances, nonsuspicious lung nodules were erroneously reported as distant metastases, a significant misclassification in the context of cancer staging.

Absence of Additional Irrelevant Findings

It is noteworthy that GPT-4 did not introduce any extraneous findings that were not originally present in the free-text reports. ►**Fig. 2** shows the graph summarizing the number of different mistakes noted in the structured reports.

Performance of GPT-4 for Generating Structured Reports Using the Revised Template

Using the revised template addressing the various areas of discrepancies noted in the initial phase of the study, GPT-4 successfully generated structured reports from 100 different HNCa CT free-text reports. In this second iteration, there were no instances of repeated mistakes or the introduction of new errors. ►**Fig. 3** shows an example of a report with missing information using the initial structured template that was correctly captured by GPT-4 in another report using the revised template. Examples of original free-text CT reports along with the GPT-4 generated structured reports are shown in ►**Supplementary Appendix 3** and ►**Supplementary**

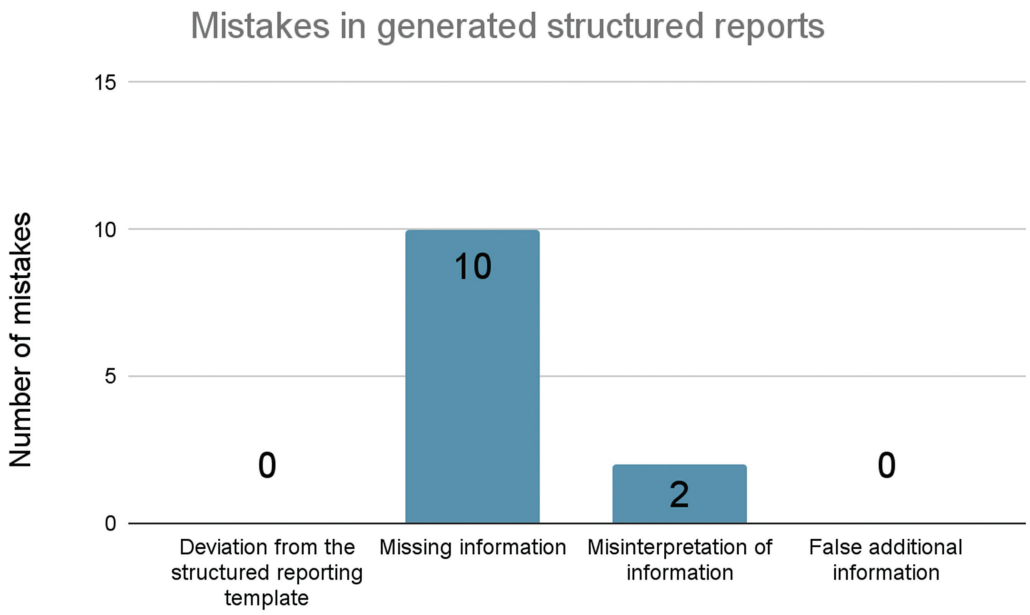

Fig. 2 Bar chart summarizing the results of the assessment for various types of mistakes in the structured reports generated by Generative Pre-trained Transformer 4 (GPT-4) using the initial structured report template.

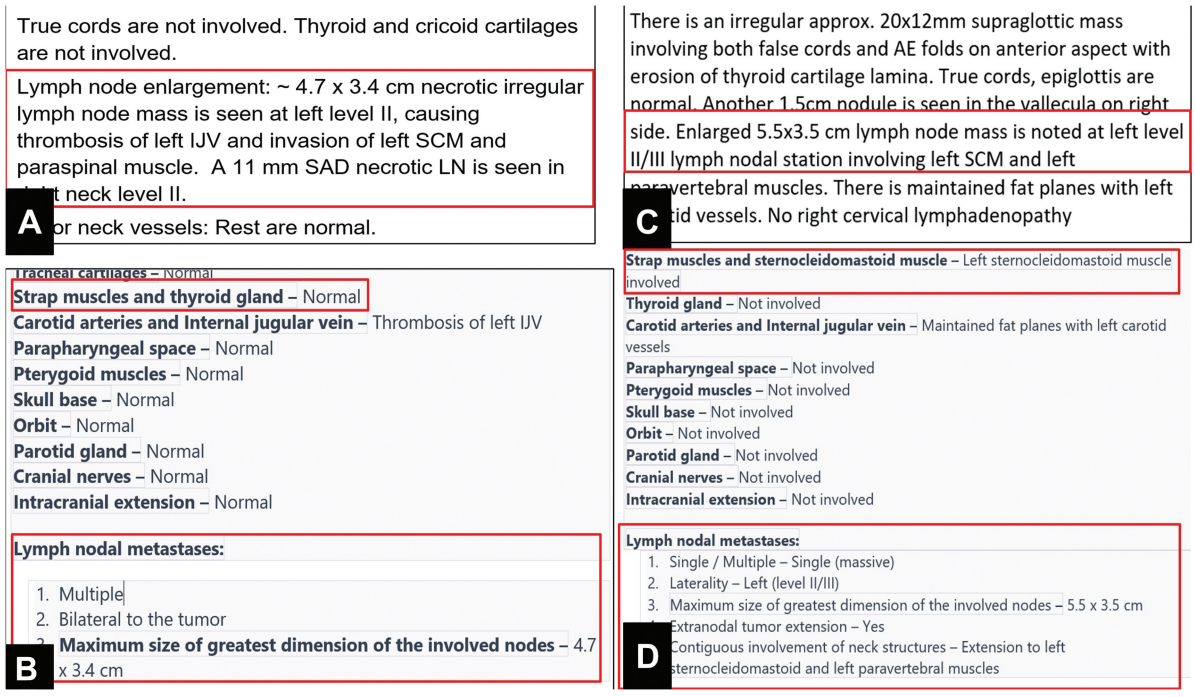

Fig. 3 Example showing an instance of missing information in the initial reporting template corrected in the revised template by the Generative Pre-trained Transformer 4 (GPT-4). (**A**) An excerpt from the conventional computed tomography (CT) report of a patient with carcinoma pyriform sinus notes the involvement of left sternocleidomastoid muscle by a lymph nodal mass (*rectangle*). (**B**) Using the initial reporting template, GPT-4 fails to note the involvement of sternocleidomastoid muscle in the option for neck muscles (*upper rectangle*) and also does not indicate the contiguous involvement of sternocleidomastoid muscle and paraspinal muscles by the nodal mass (*lower rectangle*). (**C**) An excerpt from the conventional CT report of another patient with carcinoma supraglottis notes the involvement of left sternocleidomastoid muscle by lymph nodal mass (*rectangle*). (**D**) Structured report generated by GPT-4 using the revised template correctly indicates the involvement of the sternocleidomastoid muscle as well as the presence of extranodal extension and contiguous involvement of the neck structures by nodal mass (*upper and lower rectangle*, respectively).

Appendix 4 (available in the online version). Detailed instructions given to the GPT-4 and snapshots of prompt and output of the GPT-4 are shown in ►**Supplementary Appendix 5** and ►**Supplementary Appendix 6** (available in the online version), respectively.

Discussion

In this study, we demonstrate the feasibility of using LLMs for converting free-text CT reports of HNCa patients into structured formats. The AI model (GPT-4) successfully adhered to a comprehensive structured reporting template, which was specifically developed for HNCa imaging. Although initial iterations revealed areas of missing or misinterpreted information, subsequent refinements to the reporting template resulted in accurate and complete structured reports for another set of HNCa CT reports, without any recurrent errors.

NLP is a broad field within AI that focuses on the interaction between computers and humans using natural language, employing techniques such as language translation and sentiment analysis. LLMs are a subset of NLP that specialize in generating and understanding text by training on extensive text datasets using deep learning techniques like the transformer architecture. While NLP encompasses a wide range of applications from speech recognition to chatbot development, LLMs are particularly adept at producing coherent, contextually appropriate text for sophisticated tasks like automated content generation and summarization.

The advent of LLMs in radiology signifies a transformative step in enhancing radiological reporting and patient engagement. These advanced AI models, adept at processing and understanding human language, have demonstrated their potential to automate crucial aspects of radiology reporting, such as the generation of clinical history, impressions, and layperson reports.[9–14] Previous studies have underscored the efficacy of GPT-4 and similar LLMs in generating structured radiology reports. These studies have shown that LLMs can reduce radiologists' cognitive load and improve reporting efficiency.[15–17] However, the complexity of HNCa reports, with intricate details on anatomical sites and subsites, poses a greater challenge compared with the general radiology reports evaluated in these studies. HNCa reports require nuanced understanding and precise documentation, a task that demands an advanced level of AI sophistication. In this study, we preferred GPT-4 over other LLMs as its user-friendly interface (Web site: https://chat.openai.com/) allows for easily setting up the instructions and prompts for the LLM including the desired template to be followed for structured reports, with only minimal programming skills.

Structured radiology reporting offers numerous benefits.[1–3] It ensures that all relevant clinical information is systematically documented, aiding in accurate staging and treatment planning. In the context of HNCa, site-wise detailed structured radiology reporting templates have been proposed previously in the literature.[18] In cases of posttreatment HNCa with follow-up scans, structured reporting can further support the integration of Neck Imaging Reporting and Data System (NI-RADS). Formulated by the American College of Radiology, NI-RADS is a standardized reporting template for imaging findings in patients with treated HNCa, which provides uniform lexicon to improve communication with the physicians with clear management recommendations.[19] However, the integration of structured reports into clinical practice remains limited despite these benefits.[1,6,7] Using LLMs for structured report generation from conventional free-text reports can potentially enhance their acceptance into radiologists' practice by making this process feasible with minimal interference in the radiologists' workflow and scan interpretation; it can also handle complex cases more efficiently, thus saving time and labor. Thus, LLMs can address many of the perceived cons of structured radiology reporting, hence enhancing their clinical utility.

Although utmost care was taken in the present study to anonymize any identifiable patient information before sharing the data with third-party servers, the potential data security threat can restrict the integration of the LLMs in clinical practice. In this regard, locally developed smaller language models fine-tuned for the domain-specific tasks present a viable solution. However, the widespread deployment of such language models faces significant challenges, especially due to hardware restrictions. The training and operation of language models require substantial computational resources, including advanced graphic processing units (GPUs) and large memory capacities, which can be cost-prohibitive. Additionally, the energy consumption associated with these computational demands poses sustainability concerns, further complicating their large-scale implementation.

There were a few limitations of our study. A single-reader assessment of the generated reports and inclusion of radiology reports from a single radiology department could have introduced bias due to restricted variability in the writing style and phrases used in radiology reports. We could not test the potential for integration of NI-RADS into LLM-generated structured reports in our study due to lack of posttreatment data in the collected HNCa reports.

Conclusion

In conclusion, this study demonstrates the feasibility of LLMs in transforming complex free-text radiology reports of HNCa into structured formats. The application of LLMs in radiology, particularly for complex cases like HNCa, represents a significant advancement in medical AI. However, the successful harnessing of this technology hinges on a balanced approach that combines the strengths of AI with the critical insights of radiology professionals.

Ethical Approval

Ethical approval was obtained from the institutional review board. Patient consent was not applicable for this study and was waived off by the ethics committee.

Funding

We acknowledge the support from the Ministry of Education, Government of India, Central Project Management Unit, IIT Jammu with grant sanction number IITJMU/CPMU-AI/2024/0002.

Conflict of Interest
None declared.

References

1 Marcovici PA, Taylor GA. Journal club: structured radiology reports are more complete and more effective than unstructured reports. Am J Roentgenol 2014;203(06):1265–1271
2 European Society of Radiology (ESR) ESR paper on structured reporting in radiology. Insights Imaging 2018;9(01):1–7
3 European Society of Radiology (ESR) ESR paper on structured reporting in radiology-update 2023. Insights Imaging 2023;14(01):199
4 Goldberg-Stein S, Chernyak V. Adding value in radiology reporting. J Am Coll Radiol 2019;16(9, Pt B):1292–1298
5 Pinto Dos Santos D, Baeßler B. Big data, artificial intelligence, and structured reporting. Eur Radiol Exp 2018;2(01):42
6 Shea LAG, Towbin AJ. The state of structured reporting: the nuance of standardized language. Pediatr Radiol 2019;49(04):500–508
7 Weiss DL, Langlotz CP. Structured reporting: patient care enhancement or productivity nightmare? Radiology 2008;249(03): 739–747
8 Harris D, Yousem DM, Krupinski EA, Motaghi M. Eye-tracking differences between free text and template radiology reports: a pilot study. J Med Imaging (Bellingham) 2023;10(suppl 1):S11902
9 Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. AJR Am J Roentgenol 2023;221(03):373–376
10 Lyu Q, Tan J, Zapadka ME, et al. 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(01):9
11 Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309(02):e232561
12 Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023;307(05):e230987
13 Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023;307(04): e230424
14 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(05):e230922
15 Spandorfer A, Branch C, Sharma P, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. Eur Radiol Exp 2019;3(01):37
16 Moezzi SAR, Ghaedi A, Rahmanian M, Mousavi SZ, Sami A. Application of deep learning in generating structured radiology reports: a transformer-based technique. J Digit Imaging 2023;36(01):80–90
17 Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023;307 (04):e230725
18 Mahajan A, Agarwal U, Gupta A, et al. Synoptic reporting in head and neck cancers—Head and Neck Cancer Imaging Reporting and Data Systems (HN-CIRADS): The journey ahead for standardization of imaging in head and neck cancer staging. Cancer Res Stat Treat 2022;5(02):322
19 Strauss SB, Aiken AH, Lantos JE, Phillips CD. Best practices: application of NI-RADS for posttreatment surveillance imaging of head and neck cancer. Am J Roentgenol 2021;216(06):1438–1451

