Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-025-01523-5

Cross-Institutional Evaluation of Large Language Models for Radiology Diagnosis Extraction: A Prompt-Engineering Perspective

Mana Moassefi[1] · Sina Houshmand[2] · Shahriar Faghani[1] · Peter D. Chang[3,4] · Shawn H. Sun[3] · Bardia Khosravi[1] · Aakash G. Triphati[8] · Ghulam Rasool[8] · Neil K. Bhatia[5] · Les Folio[8] · Katherine P. Andriole[5] · Judy W. Gichoya[6,7] · Bradley J. Erickson[1]

Received: 15 January 2025 / Revised: 14 April 2025 / Accepted: 23 April 2025

Abstract

The rapid evolution of large language models (LLMs) offers promising opportunities for radiology report annotation, aiding in determining the presence of specific findings. This study evaluates the effectiveness of a human-optimized prompt in labeling radiology reports across multiple institutions using LLMs. Six distinct institutions collected 500 radiology reports: 100 in each of 5 categories. A standardized Python script was distributed to participating sites, allowing the use of one common locally executed LLM with a standard human-optimized prompt. The script executed the LLM's analysis for each report and compared predictions to reference labels provided by local investigators. Models' performance using accuracy was calculated, and results were aggregated centrally. The human-optimized prompt demonstrated high consistency across sites and pathologies. Preliminary analysis indicates significant agreement between the LLM's outputs and investigator-provided reference across multiple institutions. At one site, eight LLMs were systematically compared, with Llama 3.1 70b achieving the highest performance in accurately identifying the specified findings. Comparable performance with Llama 3.1 70b was observed at two additional centers, demonstrating the model's robust adaptability to variations in report structures and institutional practices. Our findings illustrate the potential of optimized prompt engineering in leveraging LLMs for cross-institutional radiology report labeling. This approach is straightforward while maintaining high accuracy and adaptability. Future work will explore model robustness to diverse report structures and further refine prompts to improve generalizability.

Keywords LLM · Radiology · Prompt-engineering · Multi-institutional

✉ Bradley J. Erickson
BJE@Mayo.edu

[1] Mayo Clinic Artificial Intelligence Lab, Department of Radiology, Mayo Clinic, 200 1st Street, S.W., Rochester, MN 55905, USA

[2] Department of Radiology, University of California San Francisco, San Francisco, CA, USA

[3] Departments of Radiological Sciences and Computer Science, University of California, Irvine, CA, USA

[4] The Center for Artificial Intelligence in Diagnostic Medicine (CAIDM), University of California, Irvine, CA, USA

[5] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[6] Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA, USA

[7] Healthcare AI Innovation and Translational Informatics (HITI) Lab, Emory University School of Medicine, Atlanta, GA, USA

[8] Moffitt Cancer Center, Tampa, FL, USA

Abbreviations

LLMs	Large language models
GPT	Generative Pre-Trained Transformers version
LLaMA	Large Language Model Meta AI
CSV	Comma-separated values
UCSF	University of California San Francisco
MGH	Massachusetts General Hospital
UCI	University of California Irvine
SAH	Subarachnoid hemorrhage
CSp	Cervical spine fracture

Introduction

Radiology reports play a crucial role in synthesizing and clarifying complex health information. They provide essential insights that aid in accurate diagnosis and treatment planning for patients, serving as a bridge between intricate imaging data and informed medical decisions thereby

Published online: 08 May 2025

enhancing research, education, and quality improvement efforts [1]. Many efforts and publications have highlighted the significance of structured reporting in facilitating improved accuracy, consistency, and clarity in radiology practice [2–4]. One example is the Annotation and Image Markup (AIM) project, supported by the National Institutes of Health's (NIH) National Cancer Institute's (NCI) Cancer Bioinformatics Grid (caBIG™) initiative, which introduced a standardized framework for embedding annotations into clinical images, creating both human- and machine-readable data and laying the foundation for standardized practices in clinical and research communities [5]. Despite its potential, structured reporting has not gained widespread adoption, with many radiology reports remaining in freestyle formats [6, 7]. One of the primary barriers to adopting structured reporting has been its perceived disruption to the traditional speech-based workflow. This reluctance is understandable, especially given the escalating workloads faced by radiologists and the lack of additional time or resources to support such transitions [6–8].

In recent years, the literature has increasingly explored using traditional natural language processing models, large language models (LLMs), and other commercial AI-based solutions to tackle the challenges of transitioning to structured reporting [9–15]. The initial approaches to understanding medical documents primarily utilized natural language processing (NLP) methods, such as NLTK and SpaCy. These methods focused on identifying keywords as entities and extracting answers based on the presence or absence of those terms. The introduction of bidirectional encoders, specifically BERT [16], BioBERT [17], and MedBERT [18], transformed this landscape by enabling models to process text with minimal preprocessing. A key shortcoming of this approach is that words of negation may not be easy to identify, resulting in poor performance. Instead of relying solely on keyword identification, LLMs extract computational relationships between words based on extensive training on large corpora. This capability allows LLMs to effectively extract specific content and answer specific questions from medical documents. Consequently, we believe that LLMs can be instrumental in identifying the presence of specific findings or diagnoses in radiology reports. LLMs have demonstrated the capability to transform unstructured reports into structured formats, align free-text elements with corresponding sections in structured report templates, and even convert text into data formats suitable for mining [9, 10].

A variety of proprietary and open-source LLMs are now available, including OpenAI's Generative Pre-Trained Transformer (GPT) and Meta's Large Language Model Meta AI (LLaMA), which have been trained on expansive and diverse datasets spanning numerous domains, topics, and languages. This extensive training provides them with a broad knowledge base, enabling high proficiency across various subjects without requiring domain-specific fine-tuning. Notably, these models demonstrate impressive zero-shot learning capabilities, allowing them to understand and perform tasks for which they have not been explicitly trained [19]. LLaMA models are typically smaller in size than the latest GPT (GPT-4) while maintaining strong performance, especially for tasks that do not require the extensive resources that GPT-4 utilizes.

The primary goal of this project was to evaluate how effectively LLMs can identify the presence or absence of a diagnosis or conclusion within radiology reports. While it is well known that radiology groups—and individual radiologists—exhibit unique reporting styles, there is hope that LLMs may be less affected by these variations compared to traditional NLP. However, the sensitivity of LLMs to such stylistic differences among individuals or different radiology subspecialty reports remains an open question and warrants further investigation.

To explore this, we compared radiology reports from multiple centers and also assessed the performance of various models within a single center to determine whether one outperformed the others or if they demonstrated similar levels of reliability.

Methods

This multi-institutional study involved six participating institutions: Mayo Clinic, University of California, San Francisco (UCSF), Brigham and Women's Hospital and Massachusetts General Hospital (Mass General Brigham (MGB)), University of California, Irvine (UCI), Moffit Cancer Center, and Emory University. Each institution obtained approval or exemption from their respective Institutional Review Boards before initiating the study.

Each site collected 100 radiology reports for each of the following five diagnostic categories: liver metastases on abdominal CT, subarachnoid hemorrhage(SAH) on brain CT, pneumonia on chest radiograph, cervical spine fracture(CSp) on CT, and glioma progression on brain MRI. For each diagnostic category, 50 positive reports (indicating the presence of the finding) and 50 negative reports (indicating the absence of the finding) were included, resulting in a total of 500 reports per site. The diagnostic categories were determined through a manual review of the reports conducted by radiologists at each participating center. The diagnostic categories were determined through a manual review of the reports conducted by radiologists at each participating center.

To optimize the performance of the LLM in extracting relevant diagnostic findings, a structured prompt engineering process was conducted at Mayo Clinic before deployment across all participating institutions. This process involved three iterative refinements using a separate set of 30 reports

for each diagnostic category. Initially, a baseline prompt was designed and applied to the reports, and the generated outputs were systematically analyzed for errors and inconsistencies. Based on these observations, the prompt was revised to improve clarity, specificity, and alignment with radiological terminology. This refinement process was repeated two more times, each iteration incorporating insights from error patterns observed in previous rounds. The final optimized prompt was then adopted as the standardized prompt for all institutions, ensuring consistency in how the LLM processed radiology reports across sites.

A Python script was developed, that would load the 500 reports, identify the category of each report, and then use a category-specific prompt to evaluate the presence of that finding. Figure 1 shows a portion of the script, specifically the prompt for each category. All sites were asked to use the LLaMA3.1 70 billion parameter ("llama3.1 70b"). LLMs were hosted locally to ensure data privacy and compliance with institutional regulations and minimize potential latency issues during model inference.

The LLMs were deployed and executed locally at each participating site using the Ollama platform (https://ollama.com/) which provides an efficient and secure framework for running advanced LLMs locally, allowing institutions to process data on-site without relying on external servers or cloud-based solutions.

The script generated predictions for all 500 reports at each site, recording the results in a comma-separated values (CSV) file. The CSV files included key information including the category of finding, the reference label, the predicted label, and the general information like the model used and execution time. Each site then returned its respective CSV file(s) to a central repository, where the results were compiled and analyzed. The Python script instructed the LLM to provide a single-word response (e.g., "Yes" or "No"), and in cases where the LLM did not answer with that target word, it was considered incorrect.

Results

The initial analysis evaluates the performance of Llama3.1 70B across all six sites. Variations in performance likely reflect differences in reporting styles and the sensitivity of the LLM to these stylistic variations. Figure 2 presents the results for Llama3.1, encompassing all 500 reports and stratified by category. Overall, the model demonstrated strong performance, with certain categories achieving near-perfect accuracy. However, other categories, such as pneumonia, exhibited lower performance and greater variability, highlighting potential challenges in consistency across specific diagnostic domains.

Differences in performance were also observed among the various other LLMs tested (Fig. 3) with the data from one of the centers (Mayo Clinic). These variations are expected, as some models had significantly fewer parameters compared to the 70-billion-parameter reference LLM. The results demonstrated that the "ChatQA" model exhibited lower performance compared to the Instruct models. Overall, larger models outperformed smaller models, indicating a correlation between model size and accuracy.

```
def get_question(exam_class:str)->str:
    if exam_class.lower() == "cervical spine fracture":
        # apparently by using 'is' it considered only acute fractures. Could add 'or was' to include old fractures
        return ("Is there likely or definitely an acute fracture (displaced on non-displaced) of any part of the cervical spine including C1, C2, C3,
C4, C5, C6, C7, or of the odontoid? You should consider any part of the spine (the body, lateral mass, lamina, posterior elements, transverse process,
spinous process, or osteophytes as part of the spine) Answer using these options: ['Yes', 'No']. ")
    if exam_class.lower() == "pulmonary embolism":
        return ("Is there likely or definitely a pulmonary embolism, which may appear as a filling defect in a pulmonary artery, present? Options are:
['Yes', 'No']. If not specifically mentioned, then answer 'No'")
    if exam_class.lower() == "pneumonia":
        return ("Is there concern for pneumonia or a developing opacity in the lung? Options are: ['Yes', 'No']. If not specifically mentioned, then
answer 'No'")
    if exam_class.lower() == "liver metastases":
        return ("Is there likely or definitely 1 or more metastases to the liver (do not include other organs)? Options are: ['Yes', 'No'].  If not
specifically mentioned, then answer 'No'")
    if exam_class.lower() == "subarachnoid hemorrhage":
        return ("Is there likely or definitely subarachnoid hemorrhage (SAH) present (do not include other types of intracranial hemorrhage if
subarachnoid is not present)? Options are: ['Yes', 'No'].  If not specifically mentioned, then answer 'No'")
    if exam_class.lower() == "glioma progression":
        return ("What changes are seen in the brain tumor compared to only the most recent examination? Options are: ['Progression', 'Stable',
'Improved', 'Pseudoprogression', 'pseudoreponse']. Usually, increase in size means progression, and decrease in size means improved, but
pseudoprogression and pseudoresponse can be exceptions to this. if there is no clear change in the tumor except for post-operative changes or if tumor
status is not mentioned, then it is Stable. Only use the options provided, NOT the BT category.")
    return None
```

Fig. 1 Python function for dynamically generating prompts tailored to specific radiological diagnoses. The function ensures standardized question formatting and response options for various diagnostic categories, including cervical spine fractures, pneumonia, liver metastases, subarachnoid hemorrhage, and glioma progression. The prompts incorporate diagnostic nuances and provide predefined answer choices to maintain consistency and facilitate automated analysis across multiple sites

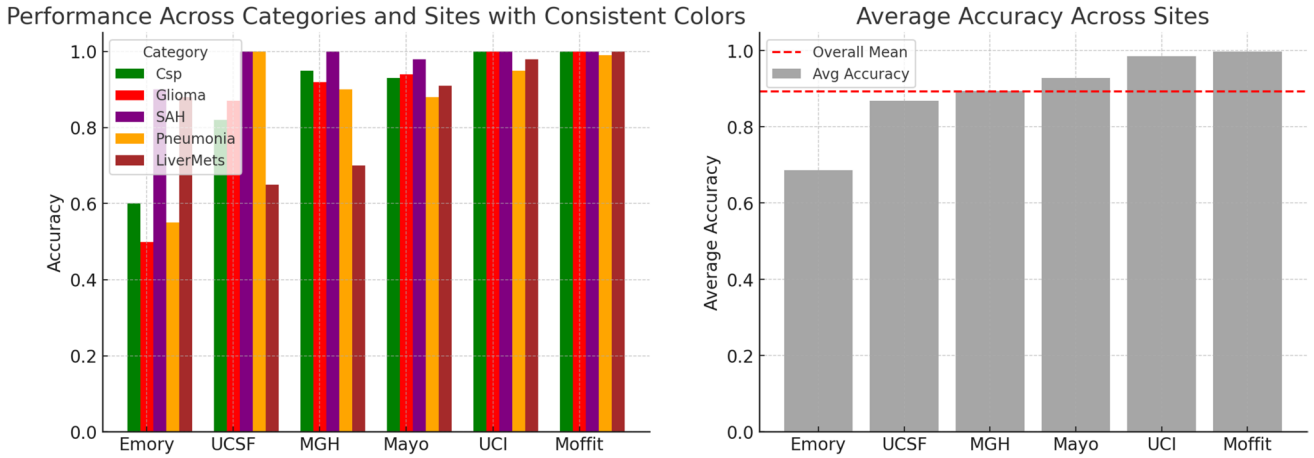

Fig. 2 The performance of the Llama3.1 70-billion-parameter model on reports from each site. The observed variability likely reflects differences in reporting styles across sites. Notably, the prompt was optimized for a single site rather than being tailored for all sites collectively (University of California San Francisco (UCSF), Brigham and Women's Hospital and Massachusetts General Hospital (Mass General Brigham (MGB)), University of California, Irvine (UCI), Subarachnoid Hemorrhage (SAH), Cervical Spine Fracture (CSp), Metastasis (Mets))

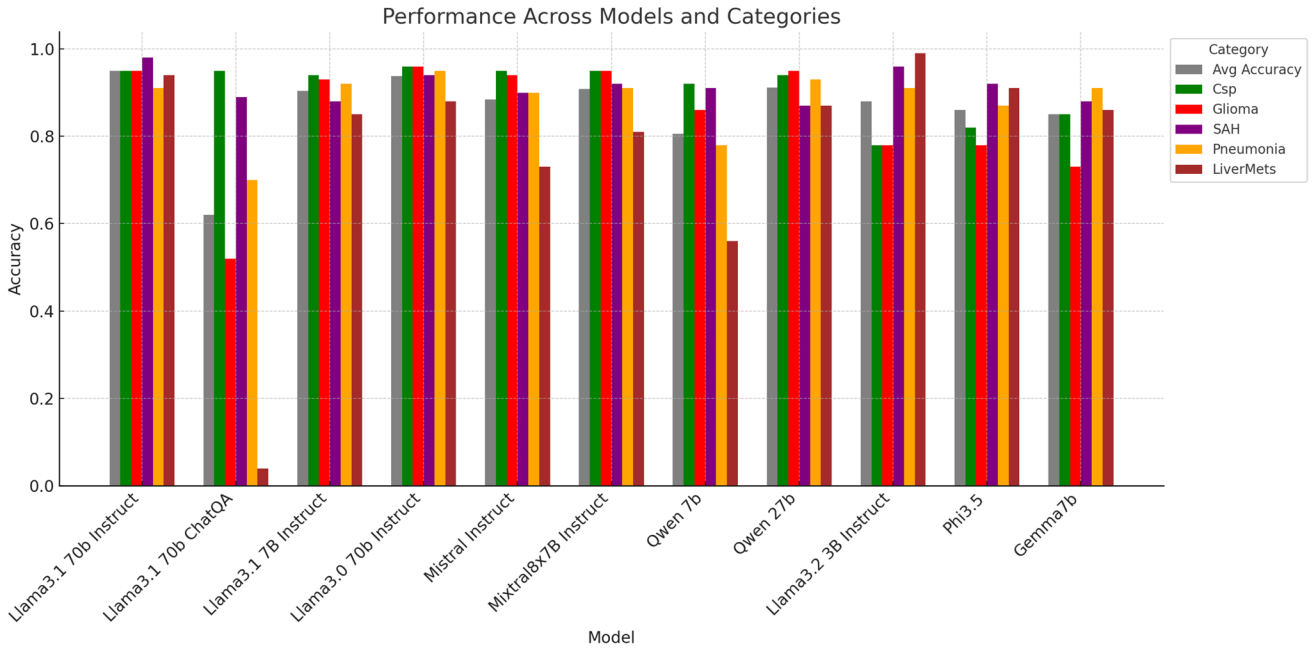

Fig. 3 The results for several different LLMs applied to one site data set (Mayo Clinic). SAH, subarachnoid hemorrhage; CSp, cervical spine fracture; Mets, metastasis; b, billion; LLaMA, Large Language Model Meta AI

Discussion

LLMs have shown remarkable capability in comprehending textual content, even in the presence of significant variability in how specific concepts are articulated. This ability is particularly valuable in the medical domain, where varying degrees of certainty, numerous synonyms, and hierarchical relationships must be accurately interpreted and understood [20]. The use of LLMs to identify key findings in radiology reports could serve important clinical and research purposes. They can quickly summarize prior exams for radiologists and clinicians [21, 22], generate patient-friendly reports with less technical jargon [23], and facilitate labeling training datasets for AI development. For research, LLMs can efficiently generate cohorts by accurately identifying findings or diagnoses within large collections of reports, streamlining the creation of extensive databases, and enhancing dataset quality for training and validation.

Our primary goal was to leverage LLMs for use cases like developing research cohorts, where inclusion and exclusion criteria are typically well-defined, such as the presence or absence of a specific finding in a radiology report. While LLMs may also have utility in broader clinical workflows, their effectiveness depends on the clarity of the question being asked. Many clinical scenarios require answers to multiple questions, necessitating a structured approach to querying LLMs for optimal performance. For instance, a simple query might determine whether a report describes "cancer," but clinical decision-making often requires further specificity—such as identifying the type of cancer, which is often less well-defined. This challenge highlights the need for well-engineered prompts and iterative refinement to ensure LLMs deliver clinically meaningful insights.

Our findings demonstrate variable performance across centers, with average accuracies ranging from 65.7% at one site to 99.8% at another. This variability is expected and likely reflects differences in reporting practices among sites. Notably, some centers achieved 100% accuracy in certain diagnostic categories. Discussions with the centers revealed that these clinical centers required structured reporting, which likely resulted in clearer wording and improved the performance of the LLM. Variations in how findings are articulated, along with differences in how radiologists express their level of certainty regarding a diagnosis, may contribute to model failures in correctly extracting information, particularly in cases from centers that do not use structured reporting.

Certain diagnostic categories proved to be significantly more challenging for the LLMs. For example, diagnosing pneumonia on chest radiographs was particularly difficult. Pneumonia often presents with air-space opacities, a finding that is not specific, as similar appearances can result from other pathologies such as edema, atelectasis, or aspiration. Consequently, many reports categorized as "positive" for pneumonia often included language suggesting it as a potential diagnosis, contingent on supportive clinical evidence (e.g., elevated white blood cell count) while others simply described the presence of an opacity. We hypothesize that this inherent ambiguity contributed to the lower performance of LLMs in this category. Additionally, variations in patient populations (e.g., hospitalized patients, acute care clinic visits, or screening settings) likely influenced the descriptions and diagnoses in reports, further impacting LLM performance.

Among the evaluated models, Llama3.1 70b Instruct emerged as the best-performing model, achieving an average accuracy of 0.95. This model, with its substantial 70 billion parameters, demonstrated consistently high accuracy across all diagnostic categories, including strong performance in more challenging domains such as pneumonia (0.91) and liver metastases (0.94). In contrast, the Llama3.1 70b ChatQA model exhibited the lowest performance, with an average accuracy of 0.62, despite having the same parameter size of 70 billion. This discrepancy underscores the impact of fine-tuning and task-specific training, as the ChatQA model was likely optimized for conversational tasks rather than structured clinical tasks like radiology report analysis. In general, the results indicate that larger models, such as those with 70 billion parameters, tend to outperform smaller models across most diagnostic categories. For instance, the smaller Llama3.1 7B Instruct model, with 7 billion parameters, achieved an average accuracy of 0.904, which, while impressive, was still below that of its larger counterpart. Similarly, models with even fewer parameters, such as Qwen 7b (7 billion parameters) and Llama3.2 3B Instruct (3 billion parameters), exhibited lower accuracies (0.806 and 0.88, respectively).

There are several important limitations to acknowledge. First, the prompt was specifically optimized for the reports from a single site. It is likely that better performance could have been achieved if reports from all sites had been available for optimization. However, this would have posed significant logistical challenges. Importantly, this also demonstrates that LLMs can still achieve good performance even without site-specific optimization. Second, the LLMs evaluated in this study are now outdated. The multi-site nature of the project required extensive coordination, and by the time results were collected, newer LLMs had been released. This limitation, however, also highlights the continuous advancement of LLMs over time, suggesting that newer models may further improve upon the findings reported here.

We also note that for a small percentage of cases, the LLMs did not respond with the required short answer (e.g., "Yes" or "No") but either provided a long explanation that usually was correct ("The report indicates the presence of an opacity that could represent pneumonia so the correct answer is 'Yes'"), or simply quoted the report ("3. Mild interval increase in the size of the brainstem glioma"). However, these were counted as incorrect because the LLM did not follow the prompt since the goal of this project was to determine if LLMs could efficiently create large databases of specific findings and so requiring an extra step of a human to review the predictions is not acceptable.

A key area for future research is confirming the robustness of our finding that structured reports significantly improve LLM performance. While our data suggest that centers utilizing structured reporting achieve higher accuracy, further studies should systematically evaluate this across multiple institutions to determine the extent of its impact. Another important direction is exploring whether a single, adaptable prompt can be effectively modified for various research applications. While prompts will likely require some customization depending on the specific research question, developing a standardized base framework could enhance efficiency and consistency in cohort generation. Additionally, future studies should investigate whether agentic AI approaches, where LLMs iteratively refine their responses based on contextual feedback, or test-time adaptation techniques, which allow

models to adjust dynamically based on new data, can further enhance performance across institutions. These strategies may help mitigate variability in report structure and terminology, improving the reliability of LLM-based information extraction in diverse clinical settings.

Conclusion

This study showed that LLMs can accurately and efficiently label large numbers of radiology reports. This is valuable for creating large research cohorts and for clinical summarization and possibly for population monitoring or AI tool monitoring. The larger LLMs tended to perform better and tended to perform better on reports that adhered to structured reporting. We anticipate that continued progress in LLM technology will result in improved performance, further increasing the value of LLMs for clinical practice and research.

Author Contribution M.M. and S.F. prepared the initial draft of the manuscript. The other authors contributed by providing reports, labels, and model results from various institutions. All authors reviewed and revised the manuscript.

Data Availability No model was trained, and the data used for inference cannot be shared at this stage.

Declarations

Competing interests The authors declare no competing interests.

References

1. Kahn CE Jr, Langlotz CP, Burnside ES, Carrino JA, Channin DS, Hovsepian DM, et al. Toward best practices in radiology reporting. Radiology. 2009;252: 852–856.
2. Bosmans JML, Neri E, Ratib O, Kahn CE Jr. Structured reporting: a fusion reactor hungry for fuel. Insights Imaging. 2015;6: 129–132.
3. Langlotz CP. Automatic structuring of radiology reports: harbinger of a second information revolution in radiology. Radiology. 2002;224: 5–7.
4. Akinci D'Antonoli T, Bluethgen C. A New Era of text mining in radiology with privacy-preserving LLMs. Radiol Artif Intell. 2024;6: e240261.
5. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL. The caBIG annotation and image Markup project. J Digit Imaging. 2010;23: 217–225.
6. Pinto Dos Santos D, Cuocolo R, Huisman M. O structured reporting, where art thou? Eur Radiol. 2024;34: 4193–4194.
7. Nobel JM, van Geel K, Robben SGF. Structured reporting in radiology: a systematic review to explore its potential. Eur Radiol. 2022;32: 2837–2854.
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: S11902.
9. Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, et al. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol. 2024;34: 3566–3574.
10. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. Radiology. 2023;307: e230725.
11. Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. Radiographics. 2001;21: 237–245.
12. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. Radiology. 2018;286: 845–852.
13. Lau WAD. Text Mining with Deep Learning for Secondary Use in Radiology. 2021. Available: https://www.proquest.com/openview/2d5b160d3b7ac791d097976d38be7510/1?cbl=18750&diss=y&pq-origsite=gscholar
14. Bhayana R. Chatbots and large language models in radiology: A practical primer for clinical and research applications. Radiology. 2024;310: e232756.
15. Le Guellec B, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an Open-Source large Language Model in extracting information from free-text radiology reports. Radiol Artif Intell. 2024;6: e230364.
16. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional Transformers for language understanding. arXiv preprint arXiv:1810.04805. Available: http://arxiv.org/abs/1810.04805. Accessed 11 Dec 2024.
17. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–1240.
18. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4: 86.
19. Kim S, Lee C-K, Kim S-S. Large language models: A guide for radiologists. Korean J Radiol. 2024;25: 126–133.
20. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620: 172–180.
21. Lee C, Vogt KA, Kumar S. Prospects for AI clinical summarization to reduce the burden of patient chart review. Front Digit Health. 2024;6: 1475092.
22. Chien A, Tang H, Jagessar B, Chang K-W, Peng N, Nael K, et al. AI-assisted summarization of radiologic reports: Evaluating GPT-3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. AJNR Am J Neuroradiol. 2024;45: 244–248.
23. Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. Sci Rep. 2024;14: 13218.

The study was conducted as an initiative of the Society of Imaging Informatics in Medicine (SIIM) Research Committee.

The corresponding author Bradley J. Erickson has following roles:.

Chair, SIIM Research Committee.

Board Member: Enquanta Inc, FlowSIGMA, Inc.

Advisor or consultant: Yunu Inc, HOPPR Inc, FindMedTech Inc.