Current Problems in Diagnostic Radiology 53 (2024) 728–737

Contents lists available at ScienceDirect

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com

Original Article

ChatGPT and assistive AI in structured radiology reporting: A systematic review

Ethan Sacoransky, BSc [a,*], Benjamin Y.M. Kwan, MD [a,b], Donald Soboleski, MD [a,b]

[a] Queen's University School of Medicine, 15 Arch St, Kingston, ON K7L 3L4, Canada
[b] Department of Diagnostic Radiology, Kingston Health Sciences Centre, Kingston, ON, Canada

ARTICLE INFO

Keywords:
ChatGPT
Artificial intelligence
Radiology
Radiologist
Large language models

ABSTRACT

Introduction: The rise of transformer-based large language models (LLMs), such as ChatGPT, has captured global attention with recent advancements in artificial intelligence (AI). ChatGPT demonstrates growing potential in structured radiology reporting—a field where AI has traditionally focused on image analysis.
Methods: A comprehensive search of MEDLINE and Embase was conducted from inception through May 2024, and primary studies discussing ChatGPT's role in structured radiology reporting were selected based on their content.
Results: Of the 268 articles screened, eight were ultimately included in this review. These articles explored various applications of ChatGPT, such as generating structured reports from unstructured reports, extracting data from free text, generating impressions from radiology findings and creating structured reports from imaging data. All studies demonstrated optimism regarding ChatGPT's potential to aid radiologists, though common critiques included data privacy concerns, reliability, medical errors, and lack of medical-specific training.
Conclusion: ChatGPT and assistive AI have significant potential to transform radiology reporting, enhancing accuracy and standardization while optimizing healthcare resources. Future developments may involve integrating dynamic few-shot prompting, ChatGPT, and Retrieval Augmented Generation (RAG) into diagnostic workflows. Continued research, development, and ethical oversight are crucial to fully realize AI's potential in radiology.

Introduction

ChatGPT (Chat Generative Pre-trained Transformer) stands as a cutting-edge large language model (LLM) developed by OpenAI, an artificial intelligence (AI) and research firm situated in San Francisco, California.[1] Introduced to the public in November 2022, ChatGPT is proficiently trained on an extensive dataset of text and code. It is capable of assisting users with various tasks such as providing answers to queries, proofreading text and crafting creative content.[2] Engaging with ChatGPT is a user-friendly experience, thanks to an intuitive chatbot feature within the online interface. Users enter text or images, coupled with prompts that guide the system, allowing ChatGPT to create specific responses suited to individual scenarios. ChatGPT boasts a user base exceeding 180 million, with the website recording 1.8 billion visits in April 2024.[3]

LLMs are advanced AI models designed to understand and generate natural language and perform prompted tasks. Built utilizing deep neural networks and trained with vast amounts of text data, these models are adept at recognizing the nuances of language and crafting responses that mimic human interaction.[4] LLMs exhibit accuracy and versatility in numerous natural language processing tasks like text summarization, sentiment analysis, language translation, analyzing grammatical structure and categorizing text. These LLMs also face numerous shortcomings, such as limitations of their training data, intrinsic model biases, knowledge gaps, data security risks, vulnerability to cyber-attacks, model "hallucinations" and ethical challenges.[4–6]

In March 2023, OpenAI released GPT-4 (Generative Pre-trained Transformer 4), a more advanced version of its predecessor GPT-3.5. GPT-4 maintains the core features of GPT-3.5 but offers enhancements, including a much larger input limit of up to 25,000 words. This version reportedly reduces error occurrences known as "hallucinations." In September 2023, OpenAI added voice and image functionalities to GPT-4, making it multimodal. It can now process text, images, and voice, performing tasks like object detection and image segmentation, and

* Corresponding author.
E-mail address: esacoransky@qmed.ca (E. Sacoransky).

https://doi.org/10.1067/j.cpradiol.2024.07.007

Available online 9 July 2024

facilitating interactive dialogues with advanced text-to-speech technology.

ChatGPT, though new to the medical field, holds untapped potential in radiology reporting—a domain where AI has predominantly focused on image analysis.[7,8] Some potential uses include its ability to convert unstructured data into structured radiology templates, generate reports from image data, create reports based on a radiologist's description of an imaging finding, and simplify reports for better patient understanding. Radiology reports are pivotal for diagnostic accuracy and influence clinical decisions and treatments.[9–11] The benefits of structured reporting are evident in studies like Brook et al.'s retrospective analysis, which demonstrated its advantages in pancreatic cancer staging, noting that structured reports include more key features and have lower inter-reporter variability.[12–14] Kabadi et al. emphasized the need for standardized terminology, suggesting that structured templates with consistent terminology can improve report quality.[14–16]

The use of ChatGPT to assist radiologists in reporting is growing. For example, in a prospective study evaluating ChatGPT's use among 286 radiology residents in India, 61.8% reported using ChatGPT while on call. Of these, 57.6% found the information provided by ChatGPT to be inaccurate, and 67.2% felt it was insufficient for diagnostic assistance. However, 85.8% were open to using it for creating report templates.[17]

The primary objective of this study was to evaluate the current utility of ChatGPT in developing structured radiology reporting templates to achieve a more standardized and thorough interpretation of imaging studies.

Methods

Comprehensive searches of electronic databases, including MEDLINE and Embase, were conducted from inception through May 2024 according to the principles of the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline.[18] Database searches were performed using the terms: "ChatGPT", "radiology", and "radiologists". A detailed search strategy is outlined in the study supplement. The search was confined to English publications. All articles were first screened, and primary studies discussing the role of ChatGPT in structured radiology reporting were selected. Any ambiguity or disagreement was resolved via consensus among all authors. The references of relevant articles were reviewed for additional sources not identified by the initial database search.

Studies investigating the use of LLMs without addressing ChatGPT were excluded, as were studies that did not focus on the use of ChatGPT for structured radiology reporting. Conference abstracts, editorials, review articles, case reports, and non-primary studies were also excluded. All authors verified the pertinence and completeness of the articles included in this review.

Pertinent data from the included studies were recorded in a predefined electronic data extraction form. This form covered study characteristics, prompting techniques, methods, outcomes, and conclusions. To ensure accuracy, the data presented in the systematic review were compared with the information in the data extraction form. Any inconsistencies were resolved through consensus among all authors.

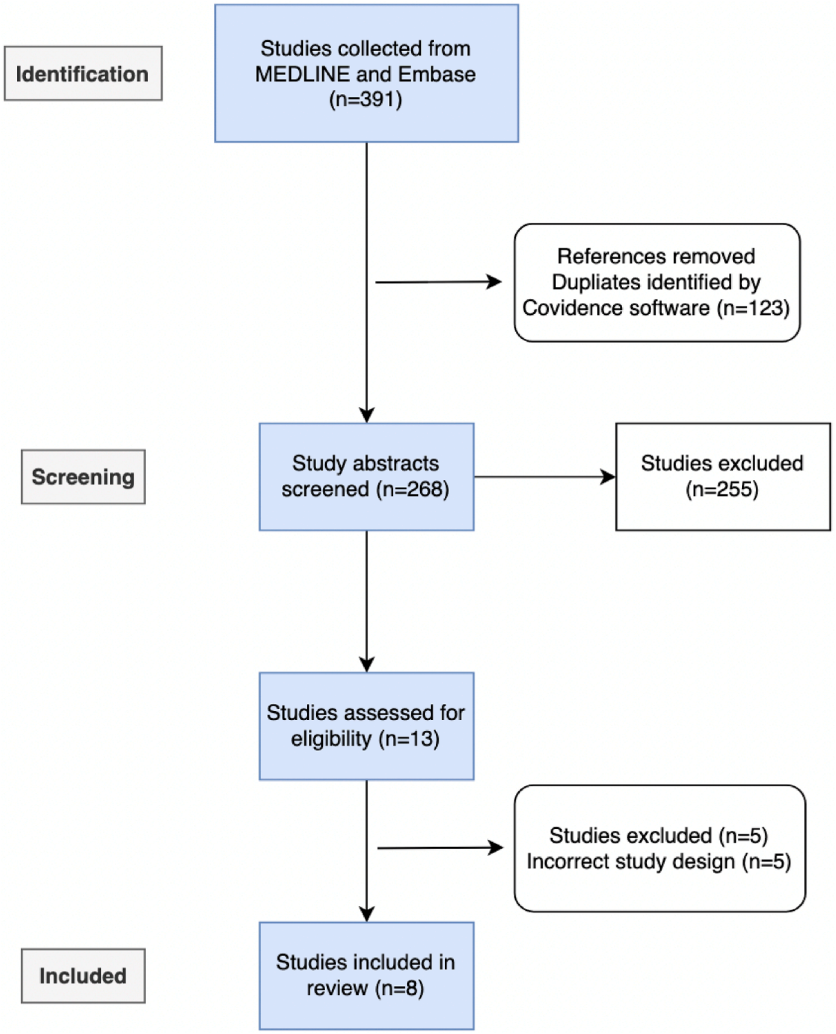

Fig. 1. PRISMA diagram of the search strategy.

Results

The literature search yielded 391 studies, of which 123 were removed due to duplication. The 268 remaining studies were individually screened, resulting in 13 articles (4.9%) being selected for full-text review. Ultimately, 8 studies (3.0%) were included in this systematic review, with publication dates ranging from March 2023 to April 2024. A PRISMA flow diagram illustrating the literature search and study selection process is presented in Fig. 1.

Table 1 offers a summary of the included articles, including study name, type, journal of publication, acceptance date, purpose, findings summary, limitations, and conclusions. The articles tested various uses of ChatGPT, including generating structured radiology reports from unstructured reports (n=3), free-text data extraction (n=2), generation of report impressions from existing findings (n=1), drafting reports from images (n=1), and structured template proposal (n=1). Some studies focused on ChatGPT's ability to structure reports for specific clinical indications such as mechanical thrombectomy in acute ischemic stroke patients, distal radius fracture classification, lung cancer characterization, and categorization of thyroid nodules. All eight studies presented an overall optimistic view of ChatGPT's potential to assist radiologists in reporting. Common critiques of ChatGPT included data privacy concerns, generalizability, reproducibility, medical errors, and a lack of optimization for medical-specific training.

In terms of prompt engineering, seven out of the eight studies employed zero-shot prompting, while one utilized reinforcement learning from a reward model based on human feedback. Seven of the eight publications used text-based inputs, whereas one used image-based inputs. One study developed a dataset of multiple input templates, four studies offered a single input template, and three studies did not provide any template as input for ChatGPT. The input data varied, encompassing CT, MRI, ultrasonography, and radiography reports across various anatomical regions. Details of the prompt engineering techniques employed in each study are outlined in Table 2.

Discussion

Our results demonstrate that ChatGPT can create structured radiology reporting templates for various pathologies, offering the potential to save radiologists valuable time while requiring minimal healthcare resources. ChatGPT-4 showed greater accuracy compared to ChatGPT-3.5, and it was effective across different imaging modalities and clinical indications.

How ChatGPT generates structured radiology reporting templates

ChatGPTs functionality is supported by an LLM, composed of a transformer neural network architecture. A neural network utilizes interconnected nodes or neurons in a layered structure like that of the brain, a process known as deep learning. Modern transformer architectures such as ChatGPT have a deep, bidirectional understanding of language, enabling them to comprehend not only the meanings of individual words but also their sequential context through positional encoding. The model's self-attention mechanism allows ChatGPT to discern the meanings of words and their relationships within a text.[29,30]

As ChatGPT structures a free-text report, it predicts the next word in the sequence in an autoregressive manner. The output is created by sequentially determining the likelihood of a specific word following another. This process is enabled by the model's training parameters, derived based on a vast corpus of text.[31] The LLM processes and understands the input text (i.e., an unstructured report) and generates the corresponding natural language output (i.e., a structured report).[32] An example of an AI system using ChatGPT in the context of structured reporting is displayed in Fig. 2.

Prompt engineering in the creation of radiology reporting templates

A prompt is a set of instructions given to an LLM that tailors, enhances, or refines its capabilities. By setting specific rules and guidelines, a prompt can significantly influence subsequent interactions with–and the outputs generated by–an LLM. A prompt establishes the context for the input and may define the expected format and content of the output. Prompt engineering refers to the process of programming LLMs by crafting prompts that guide the model towards generating the desired outputs. The quality of the outputs generated by an LLM is directly correlated with the quality of the prompts supplied by the user.[33]

In prompt engineering, using examples or templates can steer the model toward a specific format, improving output relevance and precision. Keywords focus the model's attention on the subject, and clear instructions on the output format ensure responses are appropriately structured. It is vital to create neutral prompts, free from bias or assumptions, to achieve fair and inclusive outputs. Since prompt engineering is iterative, responses often need refinement to align more closely with goals, highlighting the importance of understanding the model's capabilities and limitations.

Zero-shot, one-shot and few-shot prompting

ChatGPT demonstrates enhanced accuracy when provided with a predefined reporting template, as this gives the model a clear vision of the intended structured report.[19] In the absence of a template, the AI must rely solely on its training parameters to deduce the structure of a radiology report. Model performance can be further improved by embedding detailed instructions within the prompt regarding the expected output format. When a prompt does not include an example of the correct response (for instance, how a completed structured report should appear), this method is known as zero-shot prompting. Conversely, including a single example of a correct report in the prompt is termed one-shot prompting. The practice of adding multiple correct examples is referred to as few-shot prompting, which is currently regarded as the best practice and markedly enhances performance compared to the zero-shot or one-shot techniques.[34] Few-shot prompting is based on the principle that, akin to human learning, the AI benefits more effectively from being exposed to multiple examples. Table 3 summarizes the different prompting technique options.

Clinical accuracy of ChatGPT's responses

An important surrogate marker of ChatGPT's capability to synthesize accurate radiology reporting templates is its general knowledge of radiology, human anatomy, and pathophysiology. Several studies have been conducted in this regard, yielding mostly favorable results.

A prospective study by Barat et al. found that when ChatGPT was asked questions about common interventional radiology procedures, it provided entirely incorrect information 45% of the time, while 15% of its answers were partially accurate, and 40% were fully correct, as evaluated against standards set by two interventional radiologists.[35]

Bhayana et al. assessed ChatGPT's performance on Canadian Royal College and American Board of Radiology examination questions. ChatGPT accurately answered 69% of these questions. It performed better on lower-order thinking questions (84%) compared to higher-order thinking questions (60%).[36] A subsequent analysis showed that GPT-4 demonstrated improved performance in higher-order thinking questions compared to GPT-3.5, particularly in describing imaging findings (85% vs 61%) and applying concepts (90% vs 30%).[37]

Payne et al, evaluated GPT-4′s performance on the American College of Radiology 2022 Diagnostic Radiology In-Training Examination. GPT-4 achieved 58.5% overall accuracy, lower than the PGY-3 average (61.9%) but higher than the PGY-2 average (52.8%). Performance on image-based questions was notably poorer ($p < 0.001$) at 45.4%

Table 1
Summary of included articles.

Study Name	Article Type	Journal	Date Accepted	Study Purpose	Summary of Findings	Limitations of ChatGPT	Conclusion (Optimistic/ Cautionary)
Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study[19]	Original Research	Radiology	Mar-23	Automated the conversion of free-text radiology reports into structured templates.	GPT-4 effectively identified all key findings in the radiology reports, transforming free-text into structured reports. In every case (n=170), the model chose the most appropriate template based on the report text and its main findings.	GPT-4 requires potentially sensitive data to be shared with third parties, conflicting with privacy laws.	Optimistic
Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing[20]	Original Research	La Radiologia Medica	May-23	Compared GPT-4, ChatGPT-3.5, Perplexity and Bing in terms of knowledge on structured reporting and template proposal.	GPT-4 thoroughly elucidated the structured radiological template for a total-body CT, providing an example in tabular format. Among the four LLMs evaluated, GPT-4 was the most detailed.	LLMs were not specifically trained to generate radiology reports. Targeted training could enhance the performance and applicability of the models.	Optimistic
Evaluating GPT-4 on Impressions Generation in Radiology Reports[21]	Original Research	Radiology	Jun-23	Examined the capabilities and limitations of GPT-4 in performing zero-shot generation of radiology report impressions from pre-existing radiology report findings.	Radiologist-generated impressions were found by radiologists to be significantly ($P < .001$) more coherent, comprehensive, and factually consistent, with less medical harm than those generated by GPT-4. Conversely, referring physicians perceived GPT-4-generated impressions as more coherent and less harmful than those produced by radiologists ($P < .001$).	GPT-4-generated Impressions sometimes utilized unsupported statements, missed important information and created a certainty illusion.	Optimistic
Zero-shot information extraction from radiological reports using ChatGPT[22]	Retrospective Study	International Journal of Medical Informatics	Dec-23	Explored whether ChatGPT can extract information from radiological reports in a zero-shot manner to create structured from unstructured radiology reports.	ChatGPT can outperform baseline systems in extracting information for structured radiological reports in a zero-shot manner, with prior medical knowledge instructions boosting performance for specific tasks, albeit possibly reducing effectiveness for complex tasks.	Privacy of medical data, misunderstanding of synonyms, difficulty understanding and reasoning through complex clinical questions, inconsistency of ChatGPT outputs.	Optimistic
Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis[23]	Retrospective Study	Radiology	Mar-24	To assess whether GPT-4 and GPT-3.5 can extract data from free-text neuroradiology reports on mechanical thrombectomy in patients with acute ischemic stroke.	All free-text reports were successfully processed by GPT-4 and GPT-3.5. Of 2800 data entries, 2631 (94.0% [95% CI: 93.0, 94.8]; data points were correctly extracted by GPT-4 without the need for further postprocessing. With 1788 of 2800 correct data entries, GPT-3.5 produced fewer correct data entries than did GPT-4 (63.9% [95% CI: 62.0, 65.6]; $P < .001$). This suggests that GPT-4 could provide an alternative to retrieving these data manually.	Data extraction by GPT-4 and GPT-3.5 was only tested on a small number of reports, thus additional studies are necessary to validate the generalizability of our results. Although GPT-4 may facilitate this process and possibly improve data extraction from radiology reports, errors currently still occur and surveillance by human readers is needed.	Optimistic
Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and	Original Research	Current Problems in Diagnostic Radiology	Apr-23	The ability of ChatGPT 3.5, fine-tuned with reinforcement learning from human feedback, to draft structured radiology reports from	The study showed that musculoskeletal radiologists highly rated ChatGPT's report quality on images of distal radius fractures, demonstrating its ability to produce	Critiques mainly focused on the length of the impression sections and some limitations in handling medical terminology.	Optimistic

(*continued on next page*)

Table 1 (*continued*)

Study Name	Article Type	Journal	Date Accepted	Study Purpose	Summary of Findings	Limitations of ChatGPT	Conclusion (Optimistic/ Cautionary)
integrated AO classifier[24]				images of distal radius fractures.	competent radiology reports. This suggests significant potential for AI-driven text drafting tools in assisting radiologists.		
Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer[25]	Retrospective Study	Radiology	Aug-23	To compare the performance of GPT-3.5 and GPT-4 in data mining and labeling oncologic phenotypes from free-text CT reports on lung cancer by using user-defined prompts.	On 424 CT reports, GPT-4 surpassed GPT-3.5 in lesion parameter extraction (98.6% vs 84.0%, P < .001) and achieved 96% report accuracy, outperforming GPT-3.5s 67% (P < .001). GPT-4 also showed superior accuracy in metastatic disease identification (98.1% vs 90.3%) and oncologic progression labeling (F1 score, 0.96 vs 0.91), with higher Likert scale scores for factual correctness (4.3 vs 3.9) and accuracy (4.4 vs 3.3), and lower confabulation rates (1.7% vs 13.7%) than GPT-3.5 (all P < .001).	GPT-3.5 and GPT-4 were assessed on radiology reports from a small, homogeneous group of patients with lung cancer. Their performance may differ with report heterogeneity and lexical complexity.	Optimistic
Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography[26]	Retrospective Study	EJR	Apr-24	To assess the accuracy and reproducibility of ChatGPT in generating structured thyroid ultrasound reports from free-text reports.	On 136 ultrasound reports, ChatGPT-3.5 generated 202 satisfactory structured reports, while ChatGPT-4.0 produced only 69 (74.3% vs. 25.4%, OR = 8.490, p < 0.001). However, ChatGPT-4.0 excelled in categorizing thyroid nodules with an accuracy of 69.3% compared to ChatGPT-3.5′s 34.5% (OR = 4.282, p < 0.001) and offered better management recommendations (OR = 1.791, p < 0.001). ChatGPT-4.0 also showed higher consistency in categorizing nodules, though both versions had moderate consistency in management recommendations. These findings indicate that ChatGPT, while imperfect, is a promising tool for medical ultrasound reporting and decision support, with the potential to assist radiologists in creating structured reports that are reviewed and corrected before patient issuance.	The inter-response stability of ChatGPT remains unsatisfactory. In terms of generating management opinions, the proportion of completely incorrect opinions raised by ChatGPT-3.5 and ChatGPT-4.0 is still relatively high, and some confusing fabricated opinions may mislead junior doctors in the application process.	Optimistic

GPT = Generative Pre-trained Transformer. AI = Artificial Intelligence. RSNA = Radiological Society of North America.

compared to text-only questions (80.0%).[38]

Wagner et al. evaluated the accuracy of ChatGPT-3 in answering questions relevant to the daily tasks of radiologists distributed across eight radiological subspecialties. Findings revealed that 67% of ChatGPT-3′s responses were accurate, while 33% contained errors.[39]

In 2023, Mago and Sharma reported that ChatGPT-3′s performance in oral and maxillofacial radiology was less detailed than an expert radiologist's as it only highlighted major characteristic features. While ChatGPT-3 accurately identified radiographic landmarks, it lacked detail in pathology descriptions. The study used a 4-point Likert scale, with mean scores near 4. However, to get accurate results for the queries, the prompts needed to be very specific.[40]

In January 2024, a retrospective analysis of Radiology Diagnosis Please cases was performed by Li et al. When given clinical history and

Table 2
Prompt engineering in the included articles.

Study Name	Prompt Instructions	Input Data	Input Template	Input Format	Prompting Technique	Version of GPT
Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study	Prompt 1: "This is an unstructured report, choose the appropriate template."; Prompt 2: "Use this template to structure the free text report."	Two board-certified radiologists generated 170 fictitious CT and MRI free-text reports from various body regions and examinations.	Dataset of structured report templates based on previously published templates and the RadReport Template Library[15]	Text	Zero-shot	GPT-4
Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing	"Please provide me with an example of a structured report of a total-body CT examination; include as much detail as possible. The format must be tabular."	None	None	Text	Zero-shot	GPT-4 and ChatGPT-3.5
Evaluating GPT-4 on Impressions Generation in Radiology Reports	"Generate a new short one-line impression from the findings section using medical vocabulary."	50 free-text reports were dictated by one radiologist and three radiology residents, using chest radiographs randomly selected from the National Institutes of Health chest radiography dataset.[27]	None	Text	Zero-shot	GPT-4
Zero-shot information extraction from radiological reports using ChatGPT	"Please extract relevant information from the above report and fill in the table below."	847 pre-operative CT text reports of lung cancer patients from 2010 to 2018 at the Department of Thoracic Surgery II, Peking University Cancer Hospital.	Single template provided	Text	Zero-shot	Several versions of ChatGPT
Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis	Our prompt consisted of two parts: First, the design of the expected structured report output was explained. Second, a detailed explanation for each data point was given to ChatGPT. "Your task is as follows: Create a csv table with the following columns: Date, Localization, Side, NIHSS, ASPECTS, Lysis, Symptom Onset, Arrival, Stroke Imaging, Groin Puncture, First Intracranial Run, First Maneuver, Last Maneuver, Final Run, Number of Maneuvers, Recanalization Result, Balloon Catheter, Distal Aspiration, Stent Retriever, Extracranial Stent, Intracranial Stent, ASA, Clopidogrel, Ticagrelor, Aggrastat, Heparin, XperCT, ICH. Explanations for each column:"	Consecutive reports from patients with ischemic stroke who underwent mechanical thrombectomy between November 2022 and September 2023 were extracted from a single institution. Inclusion criteria were patient age greater than 18 years and intracranial large or medium vessel occlusion confirmed at CT or MRI with intention to treat by means of mechanical thrombectomy. Exclusion criteria were the absence of a detailed report or the absence of intracranial occlusion on angiography.	Single detailed template provided	Text	Zero-shot	ChatGPT-4 and ChatGPT-3.5
Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier	"Write a radiology report structured into exam, findings and impression which contains this exact information, do not add notes at the bottom."	Nine images of distal radius fractures were processed, with five iterations of the model run for each image.	Standard RSNA RadReport template for distal radius fracture[28]	Image	Reward model based on human feedback	Fine-tuned version of GPT-3.5
Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer	"Here is a report analysis template that consists of three steps: 1) Extract all numbers with units from the text report 2) Analyze the report and indicate whether metastasis is present (yes) or not (no) in each organ. 3) Generate an oncologic impression. Lastly, apply the template to the report I provide in the next step."	Free-text radiology reports (n=424) for lung cancer follow-up CT performed between September 2021 and March 2023 were retrieved from our institutional database.	Single template including sections on oncologic findings, metastatic site assessment and oncologic impressions.	Text	Zero-shot	GPT-3.5 and GPT-4
Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography	"Please use the following text in ultrasound medicine to generate a structured ultrasound report with patient Information, ultrasound findings, category, impression and recommendations based on ACR-TIRADS:" followed by the original radiology report in free-text.	136 free-text thyroid ultrasound reports spanning January 2023 to May 2023 from one academic research hospital.	None	Text	Zero-shot	ChatGPT-4 and ChatGPT-3.5

GPT = Generative Pre-trained Transformer. RSNA = Radiological Society of North America.

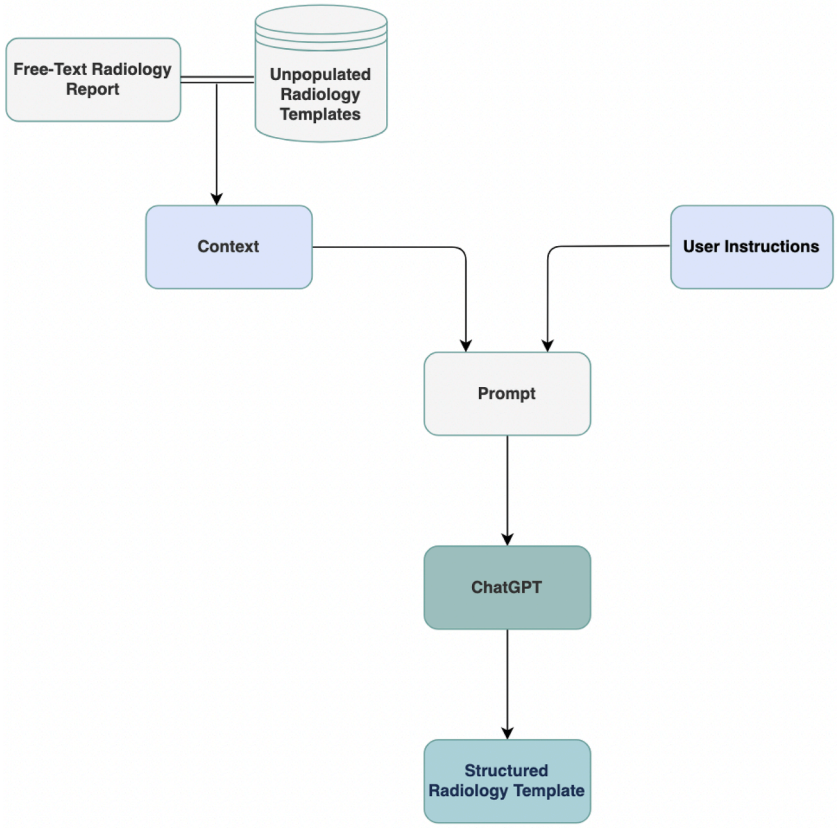

Fig. 2. ChatGPT structured reporting system diagram.

imaging findings, GPT-4 had an accuracy of 54.2%, and GPT-3.5 had an accuracy of 38.7%. They anticipate that ChatGPT accuracy on Diagnosis Please cases will continue to increase, even without radiology-specific fine-tuning.[41]

Finally, Lee et al. compared the diagnostic accuracy of the Kakao Brain Artificial Neural Network for Chest X-ray Reading (KARA-CXR) with ChatGPT. The study, involving 2,000 chest X-ray images evaluated by two radiologists, found that KARA-CXR significantly outperformed ChatGPT. KARA-CXR achieved "Acceptable" accuracy ratings of 70.50% and 68.00%, compared to ChatGPT's 40.50% and 47.00%. KARA-CXR also showed superior performance in reducing "Hallucinations," with a non-hallucination rate of 75%, significantly higher than ChatGPT's 38%. These findings emphasize the added benefit of tailoring an AI model specifically to radiology-related applications, highlighting one of the limitations of ChatGPT.[42]

Retrieval-augmented generation (RAG) and dynamic few-shot prompting

In the future, it would be beneficial for an AI system to populate the most fitting template from a database based on a radiologist's description of imaging findings. The AI's value lies in its ability to recognize the relevant template needed for classification, workup, and staging from the radiologist's free-text description. This capability could enhance reporting efficiency, especially when a final diagnosis is not yet known. However, the challenge remains in precisely matching the radiologists' findings with the correct diagnostic template, necessitating further research and development.[17]

To achieve more accurate results, we propose the use of few-shot prompting over zero-shot prompting. The work of Adams et al. showed the ability of GPT-4 to create structured reports for various body regions and examinations by selecting the most suitable template from a list.[19] This method could be appended with the use of Retrieval-Augmented Generation (RAG), a state-of-the-art AI framework that enhances LLM-generated responses by incorporating external knowledge sources. This approach improves accuracy and reduces model hallucinations by supplementing the LLM's internal data representation.[43] Specifically, it augments the LLM's context with domain-specific data.[44]

RAG operates in two phases: retrieval and content generation. During the retrieval phase, search algorithms fetch relevant information snippets from an external database or source. This external knowledge is then integrated with the user's prompt for the LLM. In the content generation phase, the LLM crafts answers based on this enriched prompt and its own data, producing responses tailored to the user's needs. While RAG is not currently integrated into the online ChatGPT interface, developers can incorporate RAG into a more sophisticated LLM application for the purpose of structured reporting.

Rau et al. integrated ChatGPT-3.5-turbo with 209 documents from the American College of Radiology appropriateness criteria guidelines using a RAG system named "LlamaIndex" to create accGPT. Tested against 50 clinical cases, accGPT's performance in creating imaging recommendations surpassed both general radiologists of varying experience levels and the generic ChatGPT-3.5 and 4.0 models, highlighting the benefits of customizing AI for specific healthcare needs with RAG.[45] We propose extending this approach to structured radiology reporting by incorporating few-shot prompting and RAG with ChatGPT. This novel approach to structured reporting, not yet explored in the radiology literature, is illustrated in Fig. 3.

The RAG framework has the capability of employing dynamic few-shot prompting, wherein the model retrieves the most relevant examples from a database and includes them as few-shot prompts. This method dynamically adds context specific to the free-text radiology report, as opposed to utilizing fixed prompt examples for every input. For example, in the case of pneumonia identified on a chest x-ray, the dynamic few-shot selection process would initially parse the unstructured report to identify key terms and findings, such as "infiltrates,"

Table 3
ChatGPT prompting techniques.

Prompting Technique	Definition	Example Prompt	Model Action
Zero-Shot	Model is given a task without any specific examples of how to perform it. The model must rely entirely on its pre-existing training parameters to generate a response.	"Given a free-text chest X-ray report, provide a structured radiology report identifying any abnormalities, normal anatomic findings, overall impressions and potential diagnoses."	The model generates a structured report based solely on its pre-existing knowledge of medical conditions, radiology terminology, and the typical content of such reports. No specific examples of radiology reports are provided to the model.
One-Shot	Providing the model with a single example to illustrate the task. This example serves as a reference for what the expected output should look like.	"Here is an example of a structured radiology report for an abdominal CT scan identifying acute appendicitis: [Example Report]. Now, given a new free-text abdominal CT report, produce a structured radiology report."	The model uses the structure, style, and type of findings detailed in the provided example report to generate a new report for the current abdominal CT scan. The example serves as a template for what findings and terminology should be included.
Few-Shot	Offering the model a few examples of how to perform the task. These examples help the model better grasp the task's nuances, variations, and the expected format of the output. Few-shot learning is especially useful for more complex tasks where a single example might not sufficiently represent the task's scope.	"Here are three examples of structured knee MRI reports, each highlighting different findings: [Example Report 1: Ligament Tear], [Example Report 2: Osteoarthritis], [Example Report 3: Meniscal Tear]. Given a new knee MRI image, create a structured radiology report."	The model analyzes the provided examples to understand the variety of potential findings in knee MRI reports and the specific language used to describe each condition. It then generates a new report for the knee MRI in question, informed by the diversity and specificity of the examples.

"consolidation," "opacities," or other imaging findings commonly associated with pneumonia. Based on these key terms, the system selects a few structured reports from a database that closely match the free-text report. This approach tailors the conversion process to the specific content of the input report, using relevant examples to guide the extraction and categorization of information. The dynamic selection of examples ensures that the structured report accurately reflects the nuances and specific findings of the original, unstructured report. This not only improves the efficiency and consistency of report generation but also enhances the utility of radiology reports for clinical decision-making and patient care.

Limitations and ethical considerations

The integration of ChatGPT into diagnostic radiology's clinical decision-making framework presents several challenges. Introducing ChatGPT into Electronic Medical Records raises concerns about patient privacy, data security, and copyright issues.[46] Additionally, there may be reluctance to utilize structured reports empirically and incorporate AI technology into radiological dictation software, which often relies on a variety of voice commands to access templates.[46–48] Furthermore, the content generated by ChatGPT can be generalized and, at times, may propagate inaccurate medical information, leading to potential errors.

In addition, the use of ChatGPT in certain radiology subspecialties may present challenges. For example, Mago and Sharma reported that ChatGPT-3′s performance in oral and maxillofacial radiology was less detailed than that of an expert radiologist, as it only highlighted the major characteristic imaging features. Since its usage by some clinicians is very specific and task-oriented, it may encounter difficulties in select medical subspecialties. Furthermore, as demonstrated by Lee et al., ChatGPT may be outperformed by other AI systems specifically trained and tailored to radiology applications. However, as shown by Fink et al., ChatGPT-4 performed very well in lung cancer lesion extraction on chest CT (98.6%) and achieved 96% report accuracy. GPT-4 also showed superior accuracy in metastatic disease identification (98.1%) and oncologic progression labeling (F1 score, 0.96). For trauma cases, Bosbach et al. demonstrated that musculoskeletal radiologists highly rated ChatGPT's report quality on images of distal radius fractures. This suggests variable performance across different subspecialties and clinical tasks, warranting future investigation.

The inherent challenges of large language models, such as misalignment with user intent, generation of fabricated information ("hallucinations"), and sourcing factual information, necessitate careful consideration when using clinically-oriented prompts.[49,50] Thus, it remains crucial to involve radiologists in clinical decision-making, providing supervision and utilizing their medical expertise. Establishing explicit guidelines for using ChatGPT and addressing errors remains essential.[51]

ChatGPT is also susceptible to biases inherent in its training data. A significant limitation of ChatGPT-4 is the outdated nature of its training data, only including information up to April 2023 and omitting the latest medical literature. This may result in data drift, a phenomenon where discrepancies between current real-world data and the model's training data lead to inaccuracies in the model. Additionally, its handling of rare pathologies may be less accurate than common ones due to imbalances in its training dataset.[50–52]

ChatGPT's effectiveness in creating radiology templates faces challenges due to its non-specialization in medical tasks and the probabilistic nature of its text generation. Enhancing its accuracy in medical contexts may require training on a custom dataset encompassing medical terminology, anatomy, and pathophysiology. ChatGPT's non-deterministic output, which relies on probabilistic token predictions, can result in variable reliability. Additionally, the lack of user control over the model makes it challenging to reproduce specific outputs. It's also important to note the rapid evolution of AI models since the release of ChatGPT. For example, GPT-3 is gradually becoming less relevant, and GPT-4 may soon be outdated with the anticipated release of GPT-5.

Lastly, the integration of AI in radiology presents ethical challenges concerning transparency, accountability, and the humanistic aspects of medicine. It's crucial to disclose the use of AI to both patients and referring physicians to maintain trust and informed decision-making. Patients need to be aware that AI assists in their diagnosis, as this can impact their perception of care quality and trust in the healthcare system. Referring physicians should also be informed to evaluate AI's performance accurately. The responsibility for diagnostic errors must be clearly defined, ensuring legal and professional accountability. The human element of medical care, including empathy and personalized communication, must be preserved despite AI's efficiency. Addressing potential biases in AI algorithms through diverse and representative training data is essential for fairness. Additionally, obtaining informed consent from patients specifically for AI use and educating them about its benefits and limitations are critical to fostering trust and informed decision-making in healthcare.

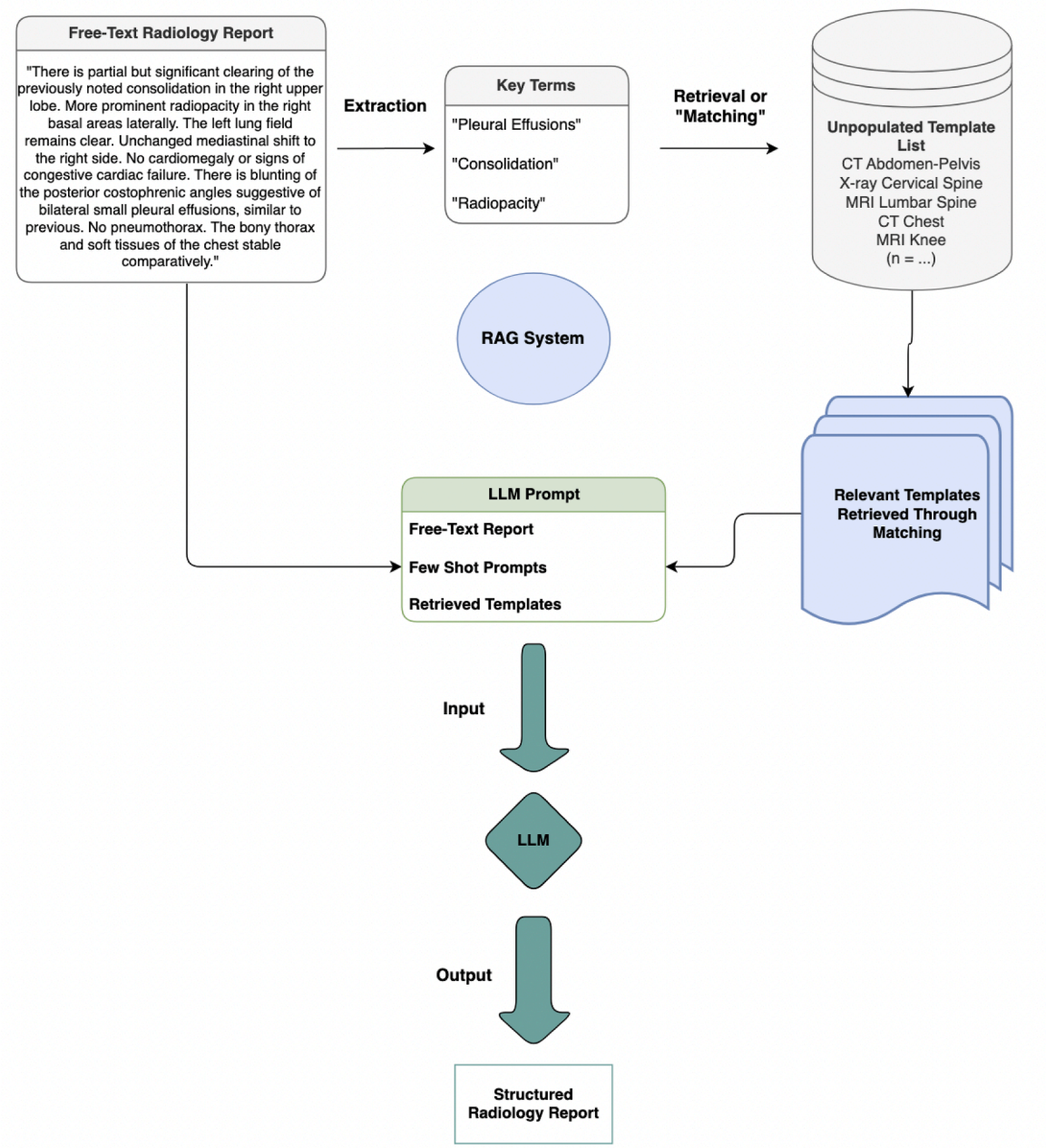

Fig. 3. Proposed method for structured radiology template generation using dynamic few-shot prompting and retrieval augmented generation (RAG).

Conclusions

ChatGPT has demonstrated its capability in creating structured radiology reporting templates for various pathologies, offering the potential to save radiologists valuable time while requiring minimal healthcare resources. Looking ahead, this technology may select the most appropriate template from a database, based on a radiologist's description of imaging findings, thereby improving the efficiency of report generation. Utilizing dynamic few-shot prompting and Retrieval-Augmented Generation (RAG) in the application of large language models (LLMs) presents opportunities to guide further research endeavors and optimize its use for disease-specific reporting. However, integrating this model architecture into daily practice and imaging systems poses challenges, particularly in user-interface design.

It is critical to acknowledge that ChatGPT's recommendations may not always align with clinical guidelines. Therefore, it is imperative for radiologists to maintain oversight and apply their clinical expertise when using artificial intelligence (AI) technology. As ChatGPT and related AI technologies continue to evolve, their ultimate role in radiology is yet to be fully determined. Addressing inherent limitations and ethical considerations is essential to ensure their secure and responsible use in healthcare settings.

Acknowledgement of grant support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Acknowledgements

We would like to acknowledge Dean Sacoransky, AI engineer, for providing valuable insight regarding AI architecture and the retrieval-augmented generation framework. We would also like to acknowledge the Queen's Department of Diagnostic Radiology Studentship (2023) for encouraging this project collaboration.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1067/j.cpradiol.2024.07.007.

References

1. Biswas SS. Role of chat GPT in public health. *Ann Biomed Eng*. 2023;51(5):868–869. https://doi.org/10.1007/s10439-023-03172-7.
2. Haupt CE, Marks M. AI-generated medical advice–GPT and beyond. *JAMA*. 2023; 329(16):1349. https://doi.org/10.1001/jama.2023.5321.
3. chat.openai.com Traffic analytics, ranking stats & tech stack. Similarweb. Accessed September 24, 2023. https://www.similarweb.com/website/chat.openai.com/.
4. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *Am J Roentgenol*. 2023;221(3):373–376. https://doi.org/10.2214/AJR.23.29198.
5. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*. 2023;307(2), e230171. https://doi.org/10.1148/radiol.230171.
6. Bavarian M., Jun H., Tezak N., et al. Efficient training of language models to fill in the middle. Published online 2022. doi:10.48550/ARXIV.2207.14255.
7. Kaka H, Zhang E, Khan N. Artificial intelligence and deep learning in neuroradiology: exploring the new frontier. *Can Assoc Radiol J*. 2021;72(1):35–44. https://doi.org/10.1177/0846537120954293.
8. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500–510. https://doi.org/10.1038/s41568-018-0016-5.
9. Lee JK, Bermel R, Bullen J, et al. Structured reporting in multiple sclerosis reduces interpretation time. *Acad Radiol*. 2021;28(12):1733–1738. https://doi.org/10.1016/j.acra.2020.08.006.
10. Pinto Dos Santos D, Hempel JM, Mildenberger P, et al. Structured reporting in clinical routine. *Fortschr Röntgenstr*. 2019;191(01):33–39. https://doi.org/10.1055/a-0636-3851.
11. Schwartz LH, Panicek DM, Berk AR, et al. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*. 2011;260(1):174–181. https://doi.org/10.1148/radiol.11101913.
12. Persigehl T, Baumhauer M, Baeßler B, et al. Structured reporting of solid and cystic pancreatic lesions in CT and MRI: consensus-based structured report templates of the german society of radiology (DRG). *Rofo*. 2020;192(07):641–656. https://doi.org/10.1055/a-1150-8217.
13. Brook OR, Brook A, Vollmer CM, et al. Structured reporting of multiphasic CT for pancreatic cancer: potential effect on staging and surgical planning. *Radiology*. 2015; 274(2):464–472. https://doi.org/10.1148/radiol.14140206.
14. Kabadi SJ, Krishnaraj A. Strategies for improving the value of the radiology report: a retrospective analysis of errors in formally over-read studies. *J Am Coll Radiol*. 2017; 14(4):459–466. https://doi.org/10.1016/j.jacr.2016.08.033.
15. RadReport. RadReport. Accessed September 23, 2023. https://www.radreport.org/.
16. Nobel JM, Van Geel K, Robben SGF. Structured reporting in radiology: a systematic review to explore its potential. *Eur Radiol*. 2022;32(4):2837–2854. https://doi.org/10.1007/s00330-021-08327-5.
17. Sethi HS, Mohapatra S, Mali C, et al. Online for on call: a study assessing the use of internet resources including ChatGPT among on-call radiology residents in India. *Indian J Radiol Imaging*. 2023;33(04):440–449. https://doi.org/10.1055/s-0043-1772465.
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
19. 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(4), e230725. https://doi.org/10.1148/radiol.230725.
20. Mallio CA, Sertorio AC, Bernetti C, et al. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol Med*. 2023;128(7):808–812. https://doi.org/10.1007/s11547-023-01651-4.
21. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology*. 2023;307(5), e231259. https://doi.org/10.1148/radiol.231259.
22. Hu D, Liu B, Zhu X, et al. Zero-shot information extraction from radiological reports using ChatGPT. *Int J Med Inform*. 2024;183, 105321. https://doi.org/10.1016/j.ijmedinf.2023.105321.
23. Lehmen NC, Dorn F, Wiest IC, et al. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. *Radiology*. 2024;311(1). https://doi.org/10.1148/radiol.232741.
24. Bosbach WA, Senge JF, Nemeth B, et al. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. *Curr Probl Diagn Radiol*. 2023, S036301882300052X. https://doi.org/10.1067/j.cpradiol.2023.04.001. Published online.
25. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*. 2023;308(3), e231362. https://doi.org/10.1148/radiol.231362.
26. Jiang H, Xia S, Yang Y, et al. Transforming free-text radiology reports into structured reports using ChatGPT: a study on thyroid ultrasonography. *Eur J Radiol*. 2024;175, 111458. https://doi.org/10.1016/j.ejrad.2024.111458.
27. Wang X, Peng Y, Lu L, et al. ChestX-Ray8: hospital-Scale Chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:3462–3471. https://doi.org/10.1109/CVPR.2017.369.
28. Weintraub MD, Hansford BG, Stilwill SE, et al. Avulsion injuries of the hand and wrist. *Radiographics*. 2020;40(1):163–180. https://doi.org/10.1148/rg.2020190085.
29. Ekin S. *Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices*; 2023. doi:10.36227/techrxiv.22683919.v1.
30. Open A.I., Achiam J., Adler S., et al. GPT-4 technical report. Published online 2023. doi:10.48550/ARXIV.2303.08774.
31. M. Fujitake, DTrOCR: decoder-only transformer for optical character recognition. Published online 2023. doi:10.48550/ARXIV.2308.15996.
32. Bang Y., Cahyawijaya S., Lee N., et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. Published online November 28, 2023. Accessed February 10, 2024. http://arxiv.org/abs/2302.04023.
33. White J., Fu Q., Hays S., et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. Published online February 21, 2023. Accessed February 10, 2024. http://arxiv.org/abs/2302.11382.
34. Ma H., Zhang C., Bian Y., et al. Fairness-guided few-shot prompting for large language models. Published online 2023. doi:10.48550/ARXIV.2303.13217.
35. Barat M, Soyer P, Dohan A. Appropriateness of recommendations provided by ChatGPT to interventional radiologists. *Can Assoc Radiol J*. 2023;74(4):758–763. https://doi.org/10.1177/08465371231170133.
36. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023; 307(5), e230582. https://doi.org/10.1148/radiol.230582.
37. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. *Radiology*. 2023;307(5), e230987. https://doi.org/10.1148/radiol.230987.
38. Payne DL, Purohit K, Morales Borrero W, et al. Performance of GPT-4 on the American college of radiology in-training examination: evaluating accuracy, model drift, and fine-tuning. *Acad Radiol*. 2024. https://doi.org/10.1016/j.acra.2024.04.006. Published online April 22.
39. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J*. 2023, 084653712311711. https://doi.org/10.1177/08465371231171125. Published online April 20.
40. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. *Cureus*. 2023. https://doi.org/10.7759/cureus.42133. Published online July 19.
41. Li D, Gupta K, Bhaduri M, et al. Comparing GPT-3.5 and GPT-4 accuracy and drift in *radiology* diagnosis please cases. *Radiology*. 2024;310(1), e232411. https://doi.org/10.1148/radiol.232411.
42. Lee KH, Lee RW, Kwon YE. Validation of a deep learning chest X-ray interpretation model: integrating large-scale AI and large language models for comparative analysis with ChatGPT. *Diagnostics*. 2023;14(1):90. https://doi.org/10.3390/diagnostics14010090 (Basel)Published 2023 Dec 30.
43. Shuster K., Poff S., Chen M., et al. Retrieval augmentation reduces hallucination in conversation. Published online 2021. doi:10.48550/ARXIV.2104.07567.
44. Lewis P., Perez E., Piktus A., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Published online 2020. doi:10.48550/ARXIV.2005.11401.
45. Rau A, Rau S, Zöller D, et al. A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology*. 2023; 308(1), e230970. https://doi.org/10.1148/radiol.230970.
46. Lee P, Bubeck S, Benefits PJ. Limits, and risks of GPT-4 as an AI Chatbot for medicine. Drazen JM, Kohane IS, Leong TY, eds. *N Engl J Med*. 2023;388(13): 1233–1239. https://doi.org/10.1056/NEJMsr2214184
47. Grewal H, Dhillon G, Monga V, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus*. 2023. https://doi.org/10.7759/cureus.40135. Published online June 8.
48. Iftikhar S, Naz I, Zahra A, et al. Report generation of lungs diseases from chest X-ray using NLP. *IJIST*. 2022;3(5):223–233. https://doi.org/10.33411/IJIST/2021030518.
49. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. *Cureus*. 2023. https://doi.org/10.7759/cureus.37589. Published online April 14.
50. Zhao W.X., Zhou K., Li J., et al. A survey of large language models. Published online September 11, 2023. Accessed September 24, 2023. http://arxiv.org/abs/2303.18223.
51. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol*. 2023, S1546144023003940. https://doi.org/10.1016/j.jacr.2023.05.003. Published online June.
52. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manag*. 2023; 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.