Hasani *et al. European Radiology* (2024) 34:3566–3574
https://doi.org/10.1007/s00330-023-10384-x

EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports

Amir M. Hasani[1], Shiva Singh[2], Aryan Zahergivar[2], Beth Ryan[3], Daniel Nethala[3], Gabriela Bravomontenegro[3], Neil Mendhiratta[3], Mark Ball[3], Faraz Farhadi[2] and Ashkan Malayeri[2*]

Abstract

Objective Radiology reporting is an essential component of clinical diagnosis and decision-making. With the advent of advanced artificial intelligence (AI) models like GPT-4 (Generative Pre-trained Transformer 4), there is growing interest in evaluating their potential for optimizing or generating radiology reports. This study aimed to compare the quality and content of radiologist-generated and GPT-4 AI-generated radiology reports.

Methods A comparative study design was employed in the study, where a total of 100 anonymized radiology reports were randomly selected and analyzed. Each report was processed by GPT-4, resulting in the generation of a corresponding AI-generated report. Quantitative and qualitative analysis techniques were utilized to assess similarities and differences between the two sets of reports.

Results The AI-generated reports showed comparable quality to radiologist-generated reports in most categories. Significant differences were observed in clarity ($p = 0.027$), ease of understanding ($p = 0.023$), and structure ($p = 0.050$), favoring the AI-generated reports. AI-generated reports were more concise, with 34.53 fewer words and 174.22 fewer characters on average, but had greater variability in sentence length. Content similarity was high, with an average Cosine Similarity of 0.85, Sequence Matcher Similarity of 0.52, BLEU Score of 0.5008, and BERTScore F1 of 0.8775.

Conclusion The results of this proof-of-concept study suggest that GPT-4 can be a reliable tool for generating standardized radiology reports, offering potential benefits such as improved efficiency, better communication, and simplified data extraction and analysis. However, limitations and ethical implications must be addressed to ensure the safe and effective implementation of this technology in clinical practice.

Clinical relevance statement The findings of this study suggest that GPT-4 (Generative Pre-trained Transformer 4), an advanced AI model, has the potential to significantly contribute to the standardization and optimization of radiology reporting, offering improved efficiency and communication in clinical practice.

Key Points

- *Large language model–generated radiology reports exhibited high content similarity and moderate structural resemblance to radiologist-generated reports.*
- *Performance metrics highlighted the strong matching of word selection and order, as well as high semantic similarity between AI and radiologist-generated reports.*

*Correspondence:
Ashkan Malayeri
Ashkan.Malayeri@nih.gov
Full list of author information is available at the end of the article

Springer

• *Large language model demonstrated potential for generating standardized radiology reports, improving efficiency and communication in clinical settings.*

Keywords Artificial intelligence, Natural language processing, Digital health, Machine learning

Introduction

In today's fast-paced medical landscape, radiologists play a vital role in clinical management, with radiology reports being their primary work product. As radiology becomes more complex, effective communication through these reports is crucial. The evolving role of radiologists in patient care highlights the need for clear communication, as diagnostic errors and inadequate communication of critical findings are leading causes of malpractice suits in the field [1].

Standardization in radiology reports corresponds to improving accuracy by matching content and structure and using unambiguous vocabulary. These standardized reports effectively convey imaging findings and recommendations to clinicians. Using standardized report templates ensures consistency and accuracy [2]. Recent advancements in artificial intelligence (AI) and natural language processing (NLP), specifically large language models (LLMs), have made standardizing radiology reports more achievable, overcoming challenges of time and expertise required for their production.

LLMs are transforming healthcare by enabling efficient data management and report processing. They have been used to been employed to generate well-structured radiology reports [3], simplify complex reports into easily understandable text [4], classify reports based on the presence or absence of specific diseases, aid in diagnostic decision-making [5, 6], provide patients with information about their examinations, outcomes, and follow-up recommendations [7], and potentially identify patterns in patient scans and reports to improve patient outcomes [8]. With the latest breakthroughs in AI and natural language generation (NLG), these models can now comprehend and generate natural language akin to human communication. Numerous studies have harnessed AI models, including deep neural networks, NLP, and transformers, to produce radiology reports from images or raw data [9–11]. Despite progress, challenges remain, such as standardizing input data quality and refining evaluation tools.

The Generative Pre-trained Transformer 4 (GPT-4), developed by OpenAI and released in March 2023, has been implemented as an option in ChatGPT Plus [12]. GPT-4 can understand and generate language consistent with human communication and has shown potential in simplifying and summarizing medical text and retrieving medical information.

In this study, we investigated GPT-4's efficacy in creating standardized radiology reports based on provided raw key findings. We conducted a comparative study comparing AI-generated reports with radiologist-generated reports, scored by four blinded readers. We also explored data collection, curation techniques, and specific prompts used with GPT-4. This investigation provides valuable insights into GPT-4's potential for enhancing the standardization and communication of radiology reports, benefiting both medical professionals and patients.

Methods

Study design and data acquisition

The study was conducted under an IRB-approved retrospective protocol. Using a comparative study design, the quality of radiology reports produced by GPT-4 and radiologists was compared. A random collection of anonymized report datasets ($n=100$) with a minimum main body length of 2000 characters was obtained from Picture Archiving and Communication Systems (PACS). These reports were generated by twelve different radiologists and spanned from 2009 to 2022. The distribution of computed tomography (CT) and magnetic resonance imaging (MRI) modalities was examined in the reports, covering chest, abdominal, and brain regions. The same set of 100 reports was used for both intervention and control groups, with the reports undergoing different processes for each group (Fig. 1).

Data curation and generation

The control group went through a manual data cleaning and restructuring process where typing errors or other non-human-generated errors were removed and restructured to mimic those of GPT-generated structure using bullet points or line breaks between sentences. This procedure did not necessitate expert intervention since the original findings remained intact. It was carried out by post-doc fellows with a 2-year background in radiology research and an international medical degree. The

(See figure on next page.)

Fig. 1 Depiction of the methodology

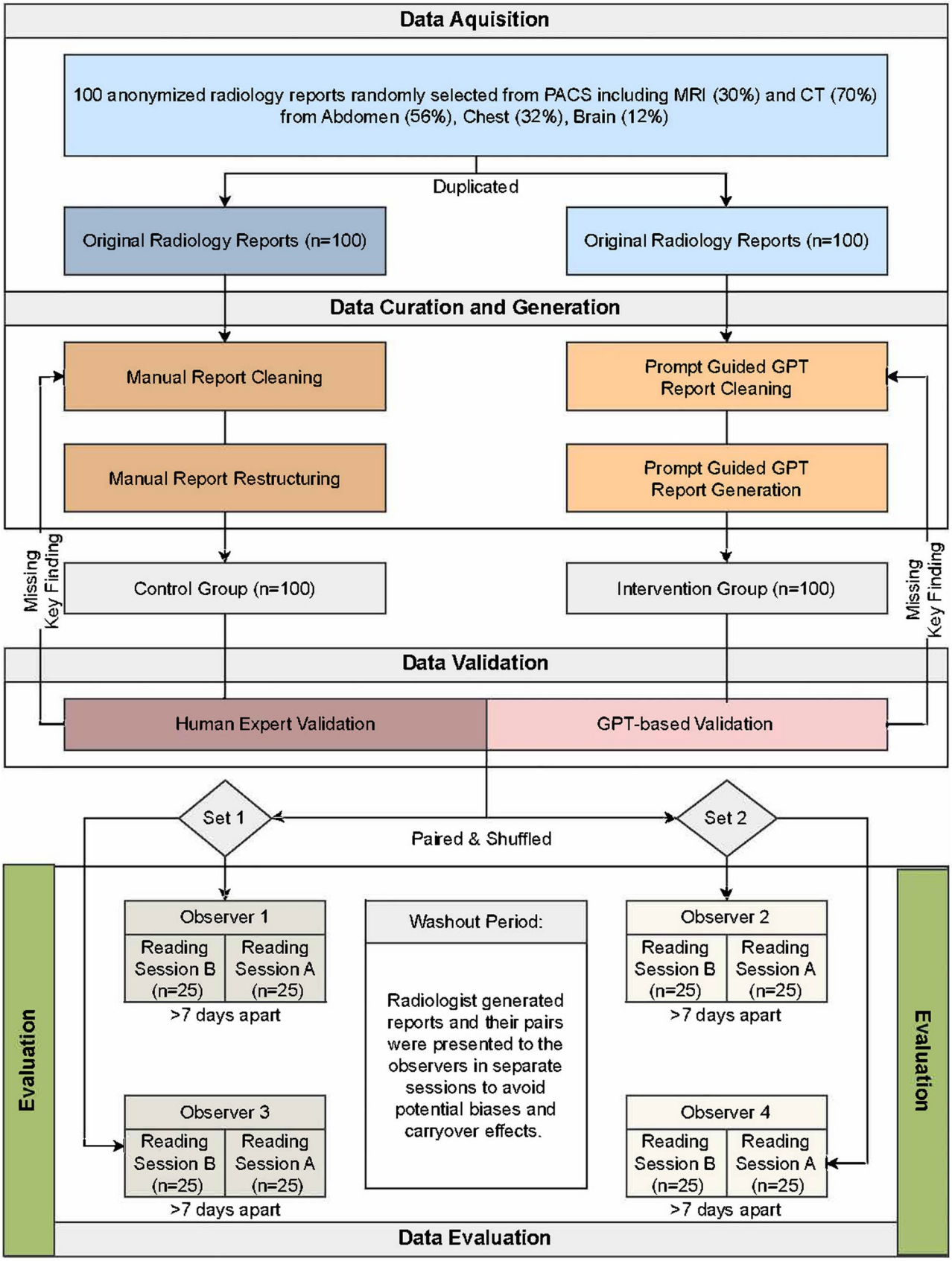

Fig. 1 (See legend on previous page.)

intervention group went through prompt-guided GPT report cleaning and prompt-guided GPT report generation following a 3-step prompt-based process shown in Table 1. The standardization process for generating high-quality radiology reports using GPT-4 involved fine-tuning settings in the OpenAI API (using Python), such as setting temperature values (0.85 or 0.9) for balanced outputs, conservative top-p values (0.9 or 0.95) for enhanced consistency, and an appropriate max tokens value (8191) for the desired report length. Careful crafting of input prompts and an iterative refinement approach were also employed to improve the quality of the generated reports. These settings were carefully chosen to prioritize coherence, consistency, and accuracy while requiring a thorough evaluation of the generated reports to ensure their reliability for clinical or research purposes.

Data validation

A two-step validation process was employed to ensure the high quality of GPT-4-generated reports. First, a retesting procedure using an independent GPT-4 prompt was performed (see Table 1) to rigorously evaluate the accuracy, relevance, and coherence of the AI-generated reports compared to the original reports, reducing the potential for randomness, errors, and misinterpretations. Each prompt was given to a new session to limit the impact of the original report on the wording of AI-generated version. Second, a post-doctorate fellow examined and validated the final GPT-4-generated reports, offering an additional layer of data verification. Both post-doctorate fellows possessed international medical degrees and had conducted a minimum of 2 years of radiology research at the National Institutes of Health. By combining these steps, adherence to the standardized format was maintained, and accuracy, relevance, and coherence were ensured, improving the overall reliability and trustworthiness of the AI-generated radiology reports. If any key information was missing from either report when compared to the original, that information was reintegrated, and the data curation and generation process was restarted for that specific report. Table 2 demonstrates a side-by-side view of radiologist-generated versus GPT-generated radiology report.

Data evaluation and bias avoidance

Four urology fellows, each with a minimum of 5 years of post-medical school experience and less than one or 2 years of experience at the National Institutes of Health National Cancer Institute, were recruited to evaluate and score the reports. They used a detailed Likert scale which included eight aspects: readability, confidence, appearance, organization, language utilization, information, relevance, and professionalism. This scale was used to guide the evaluators and ranged from 5 to 1, where 5 represented "very satisfied," 4 stood for "satisfied," 3 meant "neither satisfied nor unsatisfied," 2 indicated "unsatisfied," and 1 corresponded to "very unsatisfied." In order to minimize observer bias, the clinicians were blinded to the study's objective and whether reports were radiologist-generated or AI-generated. Each clinician evaluated 50 reports, and every report was assessed by at least two raters.

To ensure raters appraised each report independently, a wash-out period of at least 1 week was implemented between rating the human-based and AI-generated reports. This allowed for evaluating each report as a unique, standalone document. Additionally, to further minimize any potential bias, the two reading sessions for each pair of reports were scheduled with a minimum interval of 7 days between them. This time gap helped reduce the likelihood of raters recalling their previous evaluations and maintained the study's objectivity.

Table 1 List of prompts given to OpenAI GPT-4 API in order they were executed

Tasks	Prompts
Task 1: data cleaning	If there are any language, spelling errors, or grammar errors in the report: List of errors found: [list of errors] Suggestions for corrected versions: [list of corrected versions] If there are no language, spelling errors, or grammar errors in the report: [No error has been found and no further action is required.]
Task 2: key finding extraction	Extract the following information and list it in bullet point format: Key findings: [list of key findings]
Task 3: standardization	Use a standardized format appropriate for a radiology report in a clinical setting Use medical language, terminology, and abbreviations Avoid redundant and repetitive phrases and ensure that the report is easy to read and organized and includes all pertinent information
Validation prompt	Compare the radiologist-generated report A and the AI-generated report B, identify missing information in report B, and present the errors and differences in a bullet point format

Table 2 Side-by-side comparison of radiologist-generated versus GPT-4-generated radiology report

Original report	GPT-generated report
Exam: CT chest, abdomen + pelvis Reason for study: pt with BHD mutation, father with bilat kidney ca. last CT in [Date] neg, pt needs period screening pls examine and characterize thank you Clinical information: BHD Multidetector helical (5 mm, 8-channel, 44 cm dfov) images following, including multiphase abdomen images following, and abdomen images prior to vascular contrast administration (166 cm, 96.2 kg, 18 s delay, 4 cc/s, 130 cc Isovue 300) obtained without apparent complication. 2.5 mm abdomen images reconstructed from data, and enteric contrast also given Chest: several bilateral lung cavities up to about 1.5 cm length, with larger foci possibly due to bullous blebs Small spine osteophytes No evidence of pleural or pericardial effusion, bulky hilum, mediastinum, or axilla adenopathy, or lung or bone gross metastasis Abdomen, pelvis: Subcentimeter liver focus adjacent to gallbladder (e.g., series 2 image 22, series 4 image 19, 112–113) probably due to cyst Findings compatible with right extrarenal pelvis Subcentimeter periaortic and mesenteric foci compatible with small lymph nodes Minimal vascular calcification. Absent or atrophic uterus Moderate intra-abdominal fat No evidence of splenomegaly, hydronephrosis, calcification in gallbladder or kidneys, pelvis ascites, or pancreas, kidney, or bone gross untoward mass No comparison study Impression: several lung cavities possibly due to history of BHD No evidence of kidney mass to suggest carcinoma	Exam: CT chest, abdomen, and pelvis Clinical information: BHD patient with a history of bilateral kidney cancer. Last CT in [date] was negative. Patient requires periodic screening. Please examine and characterize findings Technique: multidetector helical CT (5 mm, 8-channel, 44 cm dFOV) with multiphase abdomen images following IV contrast administration (166 cm, 96.2 kg, 18 s delay, 4 cc/s, 130 cc Isovue 300) obtained without complications. 2.5 mm abdomen images reconstructed from data. Enteric contrast also given Findings: Chest: Lungs: bilateral lung cavities up to 1.5 cm in size, larger foci possibly due to bullous blebs Pleura and pericardium: no pleural or pericardial effusions Mediastinum and axilla: no bulky adenopathy or gross metastasis in the hilum, mediastinum, or axilla Bones: no osseous metastasis, small spinal osteophytes Abdomen and pelvis: Liver: subcentimeter focus adjacent to gallbladder, likely a cyst Kidneys: findings compatible with right extrarenal pelvis, no evidence of carcinoma Lymph nodes: subcentimeter periaortic and mesenteric foci, likely small lymph nodes Vasculature: minimal vascular calcification Uterus: absent or atrophic Fat: moderate intra-abdominal fat Other: no splenomegaly, hydronephrosis, gallbladder or kidney calcifications, pelvic ascites, or gross untoward mass in the pancreas, kidneys, or bones Comparison: no comparison study available Impression: 1. Several lung cavities, possibly related to the patient's history of BHD 2. No evidence of kidney mass to suggest carcinoma 3. Subcentimeter liver focus adjacent to the gallbladder, likely a cyst 4. Other findings, including small lymph nodes and minimal vascular calcification, are non-specific and likely incidental

Data analysis

Quantitative analysis

In comparing the accuracy and similarity between radiologist-generated reports and those generated by GPT-4 AI, five distinct measurements were employed: Cosine Similarity, Sequence Matcher Similarity, Different Token Count, BLEU (Bilingual Evaluation Understudy), and BERTscore (Bidirectional Encoder Representations from Transformers score). Cosine Similarity was utilized to ascertain the similarity of term frequencies within the differing reports. Sequence Matcher Similarity, derived from Python's "difflib" library, was applied to quantify the similarity between character sequences. Different Token Count was used to measure basic textual similarity, identifying shared unique words between the reports. The BLEU score was harnessed to assess the quality of AI-generated text, benchmarking it against radiologist-created reports. Finally, the BERTScore was used to compare the contextual word embeddings of the generated and reference texts, offering a nuanced evaluation of textual similarity and translation quality. A summary of these evaluation metrics used for comparing the radiology reports can be found in Table 3.

Qualitative data analysis

The data collected from the Likert scale evaluations were analyzed using descriptive and inferential statistical methods. Descriptive statistics, such as means, medians, and standard deviations, were calculated for each Likert scale category to summarize the clinicians' evaluations of both AI-generated and radiologist-generated reports. To analyze the ordinary and non-normally distributed Likert scale evaluations, two non-parametric tests, including the Wilcoxon Signed-Rank Test or the Mann–Whitney *U* Test, were employed to compare the two report types, identifying any statistically significant differences across the eight categories.

Results

Qualitative analysis

In this comparative study, a total of 100 reports were analyzed, with the breakdown showing that the abdomen

Table 3 Quantitative metrics used for comparing radiology reports

Evaluation metric	Description
Cosine Similarity	Measures the angle between two TF-IDF vectors, capturing the semantic similarity of the texts
Sequence Matcher Similarity	Calculates the proportion of matching subsequences, evaluating structural and organizational similarity
Different Token Count	Quantifies the number of word-level differences between the reports
BLEU Score	Measures the quality of text by comparing it to one or multiple reference translations
BERTScore F1	Evaluates the semantic similarity between two texts using the pre-trained BERT language model

Term Frequency–Inverse Document Frequency (TF-IDF), Bilingual Evaluation Understudy (BLEU) Score, and Bidirectional Encoder Representations from Transformers (BERT) Score F1

was the most frequently imaged body region (56%), followed by the chest (32%) and brain (12%). Furthermore, CT scans accounted for 70% of the total scans, while MRI constituted the remaining 30%.

The results of our evaluation of the radiologist-generated and GPT-generated reports across various categories are presented in Table 4. As shown, GPT-generated reports outperformed radiologist-generated reports in most categories, including clarity, confidence, ease of understanding, structure, correctness, relevancy, and adherence to medical norms. Statistically significant differences were observed in clarity ($p=0.027$), ease of understanding ($p=0.023$), and structure ($p=0.050$), indicating that the GPT-generated reports demonstrated higher satisfaction in these aspects. The stacked bar chart (Fig. 2) complements the findings from the table, visually illustrating the performance of each report type across the assessed categories. Together, these results provide insights into the performance of the GPT-4 model in generating radiology reports and highlight areas for potential improvement.

Quantitative analysis

Our analysis compared textual features of radiologist-generated and AI-generated radiology reports. Radiologist-generated reports had more words (average 338.23) and characters (average 1923.25) than AI-generated reports (average 292.49 words and 1687.61 characters). Statistical analysis revealed significant differences ($p=0.0024$ for total words, $p=0.0037$ for total characters). Both report types had similar average sentence lengths, but AI-generated reports showed greater variability in sentence length.

AI-generated reports were more concise, with 34.53 fewer words and 174.22 fewer characters on average. The differences were statistically significant ($p=0.0024$ for total words, $p=0.0037$ for total characters). Despite shorter average sentence lengths, AI-generated reports had greater diversity in sentence structure. This highlights the need for refining sentence consistency in natural language generation algorithms.

The comparison between original radiologist-authored and GPT-4 AI-generated reports revealed an average Cosine Similarity of 0.85, indicating high content similarity. Sequence Matcher Similarity averaged 0.52, denoting moderate structural resemblance, while Different Token Count analysis found an average of 33.22 distinct word-level discrepancies.

We evaluated the performance of the text generation model using BLEU Score and BERTScore F1. The model achieved a BLEU Score of 0.5008, indicating approximately half of the words and phrases in the generated text aligned with the reference text. The model also achieved a BERTScore F1 of 0.8775, suggesting a high level of

Table 4 Comparison of radiologist-generated and GPT-generated reports across various categories, including Kruskal–Wallis *H* Test results

Category	Radiologist-generated (mean ± SD)	GPT-generated (mean ± SD)	Kruskal–Wallis H Test (H)	*p* value
Clarity	3.7 ± 0.97	4.2 ± 0.82	4.92	**0.027***
Confidence	3.7 ± 0.98	4.0 ± 0.97	1.96	0.162
Ease of understanding	3.5 ± 1.05	4.0 ± 0.98	5.16	**0.023***
Structure	3.4 ± 1.18	3.8 ± 1.18	3.85	**0.049***
Correctness	3.9 ± 0.86	4.2 ± 0.84	1.63	0.201
Relevancy	3.9 ± 0.95	4.3 ± 0.70	2.67	0.102
Adherence to medical norms	3.9 ± 0.88	4.2 ± 0.87	N/A	N/A

* Statistically significant. $p < 0.05$

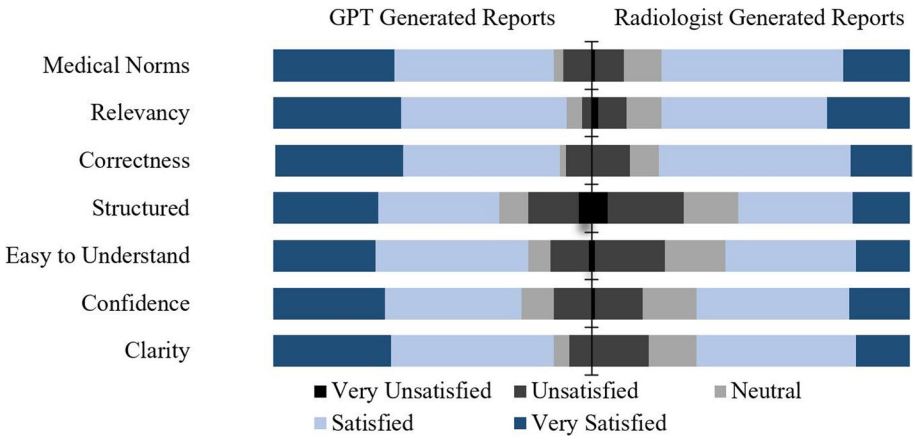

Fig. 2 The stacked bar chart illustrates the distribution of satisfaction levels for radiologist-generated and GPT-generated reports across various categories. Light blue bars represent "satisfied" responses, while dark blue bars indicate "very satisfied" responses. Notably, a substantial shift from light blue to dark blue is observed in categories with significant differences, such as clarity, confidence, ease of understanding, structure, and adherence to medical norms. This demonstrates that the GPT-generated reports led to higher levels of satisfaction in these key aspects

similarity between the generated and reference texts in capturing semantic meaning.

Discussion

In the comparative study conducted by our team, the quality of the AI-generated reports was found to be comparable to that of the radiologist-generated reports across various categories, including readability, confidence, appearance, organization, language utilization, information, pertinence, and professionalism. This efficiency in generating reports can enhance radiology practices by optimizing resources and minimizing the time and effort involved in structured reporting. Additional use of a standardized report format and terminology offered by GPT-4 could also enhance communication between radiologists, clinicians, and other healthcare professionals, thus minimizing the scope for error. Standardization also simplifies future data extraction and analysis for research and quality improvement purposes.

Our study is the first to use analysis tools like Cosine Similarity, previously applied by Mabotuwana et al [13], for comparing AI-generated and radiologist-generated reports. We found a significant average cosine similarity index of 0.85, suggesting considerable likeness between the two sets of reports. This likeness is particularly notable given that GPT-4 tends to shorten the report text. In our study, AI-generated reports had 34.53 fewer words than radiologist-generated reports, which might have affected the sequence matching score of 0.52, as observed in the findings of Lyu et al [14]. Despite the text shortening, the BLEU score, a common precision metric, was employed to further compare generated sentences, showing promising results. Studies by Jing et al [15] and Alfarghaly et al [9] achieved BLEU scores of 0.247 and 0.111, respectively. In our study, the comparison between GPT-4 AI-generated and radiologist-generated reports yielded a high BLEU score of 0.5008, indicating substantial matching of word selection and order in both sets of reports, further emphasizing the potential of GPT-4 in radiology reporting.

BERT (Bidirectional Encoder Representations from Transformer) is a transformer-based NLP model which has been used in the past for searching entity recognition in texts. Pre-trained BERT models have been employed to search for matching clinical keywords (e.g., endotracheal tube or ET tube or ETT) in labeled reports in multiple studies and classify them based on their presence or absence [16–20]. One such study was done by Olthof et al [20], where the team labeled trauma radiology reports based on the presence or absence of injuries. Pre-trained BERT models were used for the classification of the same reports which achieved an F1 score of 0.95 for simple reports and 0.83 for complex reports [20]. Another study done by Li et al [18] on classifying actionable radiology reports of tinnitus patients using BERT in-domain pre-trained model demonstrated the highest F1-score of 0.84. Our study yielded a BERT F1 score of 0.8775, demonstrating high semantic similarity between AI-generated and radiologist-generated reports. This suggests that both sets of reports convey identical connotations and do not miss any pertinent clinical information.

One of the most significant limitations of GPT-4 is its tendency to "hallucinate" inaccurate responses when the required information is absent from its training data. This can generate seemingly plausible but factually incorrect answers, which might mislead users [21].

Another challenge in GPT-4 is its restricted access, as it often necessitates sharing potentially sensitive or private data with external parties. This may lead to violations of privacy regulations, especially in industries where confidentiality is of utmost importance [12]. Furthermore, GPT-4's training data only includes information up to September 2021. This limitation renders the model incapable of offering current and up-to-date information on evolving topics, trends, or recent events. Given these limitations, users must exercise extreme caution when relying on GPT-4 outputs. This is particularly important in high-stakes situations, such as medical contexts, where misinformation or inaccuracies could have severe consequences. Users are advised to cross-check the information generated by GPT-4 with reliable sources and expert opinions to ensure its accuracy and relevance.

Before implementing GPT-4 in clinical practice, it is essential to address several limitations and ethical considerations. The limited sample size in this study raises concerns about generalizability, and further research with diverse settings and larger samples is necessary. Ensuring the accuracy, relevance, and coherence of AI-generated reports is crucial, and rigorous quality control measures and continuous monitoring are needed for safe implementation. GPT-4 is capable of generating human-like text, but it can also produce "hallucinations"—mistakes in the generated text that are semantically or syntactically plausible but are in fact incorrect or nonsensical [22]. Some researchers have identified two factors that predict much of the performance of LLMs and propose that these are major sources of hallucination in generative LLMs. The first factor is the memorization of training data, where models falsely label test samples as entailing when the hypothesis is attested in the training text, regardless of the premise [22]. The second factor is the exploitation of a corpus-based heuristic using the relative frequencies of words [22]. Another study found that LLMs, such as ChatGPT, are prone to generate hallucinations, i.e., content that conflicts with the source or cannot be verified by factual knowledge [23]. The empirical results suggest that ChatGPT is likely to generate hallucinated content on specific topics by fabricating unverifiable information (i.e., about 11.4% user queries) [23]. Moreover, existing LLMs face great challenges in recognizing hallucinations in texts [23].

The results reveal that GPT is not immune to errors. As seen in Table 2, for instance, the phrase "pt with BHD mutation, father with bilat kidney ca" was incorrectly translated by GPT as "BHD patient with a history of bilateral kidney cancer." Thus, while GPT holds promise, its current form demands cautious application in clinical settings to ensure accuracy and patient safety. Ethical considerations, such as data privacy and accountability, must also be addressed when using GPT-4, with rigorous guidelines and protocols in place to maintain patient privacy and tackle ethical concerns.

In conclusion, despite text shortening, the AI-generated reports maintained clinical and semantic relevance. GPT-4 holds potential for generating standardized radiology reports, improving efficiency and communication. However, addressing limitations and ethical implications is crucial for safe and effective implementation. Further research is essential to maximize benefits and minimize risks associated with GPT-4 or similar AI systems in radiology report generation.

Abbreviations

AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CT	Computed tomography
GPT-4	Generative Pre-trained Transformer 4
LLMs	Large language models
MRI	Magnetic resonance imaging
NLG	Natural language generation
NLP	Natural language processing
PACS	Picture Archiving and Communication Systems
SDV	Standard deviation
TF-IDF	Term Frequency–Inverse Document Frequency

Funding
This study has received funding by National Institutes of Health.

Declarations

Guarantor
The scientific guarantor of this publication is Ashkan Malayeri.

Conflict of interest
The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry
No complex statistical methods were necessary for this paper.

Informed consent
Written informed consent was not required for this study because radiology reports were anonymized.

Ethical approval
The National Institutes of Health's Institutional Review Board (IRB) approved the protocol as a retrospective study.

Study subjects or cohorts overlap
N/A.

Methodology

- Comparative quantitative and qualitative analysis

Author details
[1]Laboratory of Translation Research, National Heart Blood Lung Institute, NIH, Bethesda, MD, USA. [2]Radiology & Imaging Sciences Department, Clinical

Center, NIH, Bethesda, MD, USA. [3]Urology Oncology Branch, National Cancer Institute, NIH, Bethesda, MD, USA.

Received: 11 July 2023 Revised: 1 September 2023
Accepted: 8 September 2023 Published online: 8 November 2023

References

1. Srinivasa Babu A, Brooks ML (2015) The malpractice liability of radiology reports: minimizing the risk. Radiographics 35:547–554
2. Larson DB (2018) Strategies for implementing a standardized structured radiology reporting program. Radiographics 38:1705–1716
3. Adams LC, Truhn D, Busch F et al (2023) Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology. https://doi.org/10.1148/radiol.230725:230725
4. Jeblick K, Schachtner B, Dexl J et al (2023) ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. https://doi.org/10.1007/s00330-023-10213-1
5. Gaube S, Suresh H, Raue M et al (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 4:31
6. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD (2023) Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. https://doi.org/10.1101/2023.02.02.23285399
7. Choudhury A, Asan O (2020) Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 8:e18599
8. Aggarwal R, Sounderajah V, Martin G et al (2021) Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 4:65
9. Alfarghaly O, Khaled R, Elkorany A, Helal M, Fahmy A (2021) Automated radiology report generation using conditioned transformers. Inform Med Unlocked 24:100557
10. Monshi MMA, Poon J, Chung V (2020) Deep learning in generating radiology reports: a survey. Artif Intell Med 106:101878
11 Wiggins WF, Kitamura F, Santos I, Prevedello LM (2021) Natural language processing of radiology text reports: interactive text classification. Radiol Artif Intell 3:e210035
12. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388:1233–1239
13. Mabotuwana T, Lee MC, Cohen-Solal EV (2013) An ontology-based similarity measure for biomedical data – application to radiology reports. J Biomed Inform 46:857–868
14. Lyu Q, Tan J, Zapadka ME et al (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: promising results, limitations, and potential. https://doi.org/10.48550/arXiv.2303.09038
15. Jing B, Xie P, Xing EP (2017) On the automatic generation of medical imaging reports Annual meeting of the Association for Computational Linguistics
16 Tejani AS, Ng YS, Xi Y, Fielding JR, Browning TG, Rayan JC (2022) Performance of multiple pretrained BERT models to automate and accelerate data annotation for large datasets. Radiol Artif Intell 4:e220007
17 Yan A, McAuley J, Lu X et al (2022) RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell 4:e210258
18. Li J, Lin Y, Zhao P et al (2022) Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (BERT) and in-domain pre-training (IDPT). BMC Med Inform Decis Mak 22:200
19 Nishigaki D, Suzuki Y, Wataya T et al (2023) BERT-based transfer learning in sentence-level anatomic classification of free-text radiology reports. Radiol Artif Intell 5:e220097
20. Olthof AW, Shouche P, Fennema EM et al (2021) Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Programs Biomed 208:106304
21. OpenAI (2023) GPT-4 Technical Report. Arxiv abs/2303.08774
22. Alkaissi H, McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 15:e35179
23. Li J, Cheng X, Zhao WX, Nie J-Y, Wen J-R (2023) HELMA: a large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:230511747

Publisher's Note