Published in final edited form as:
Radiology. 2023 June ; 307(5): e231259. doi:10.1148/radiol.231259.

Evaluating GPT-4 on Impressions Generation in Radiology Reports

Zhaoyi Sun, MS[*],
Department of Population Health Sciences, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Hanley Ong, MD[*],
Department of Radiology, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Patrick Kennedy, MD,
Department of Radiology, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Liyan Tang, BS,
School of Information, The University of Texas at Austin, Austin, Tex

Shirley Chen, MD,
Department of Radiology, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Jonathan Elias, MD,
Primary Care, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Eugene Lucas, MD,
Comprehensive Weight Control Center, New York–Presbyterian/Weill Cornell Medical Center, New York, NY

George Shih, MD,
Department of Radiology, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Yifan Peng, PhD
Department of Population Health Sciences, Weill Cornell Medicine, 425 E 61st St, Suite 301, New York, NY 10065

Generation of Impressions from radiology report Findings is a critical aspect of medical image analysis, assisting clinicians in making informed decisions (1). Traditionally, this process requires manual input from the interpreting radiologist, which can be time consuming and occasionally can be inconsistent with the Findings section. Fine-tuned pretrained models have shown promise in automating or proofreading this task (2); however, they often necessitate substantial training data sets, which may not always be accessible in specialized domains, such as radiology. The recent success of large language models, such

Address correspondence to Y.P. (yip4002@med.cornell.edu).

[*] Z.S. and H.O. contributed equally to this work.

Author contributions: Guarantors of integrity of entire study, **Z.S., Y.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **Z.S., P.K., S.C., G.S., Y.P.**; clinical studies, **Z.S., H.O., L.T., S.C.**; experimental studies, **Z.S., P.K., L.T., J.E., G.S., Y.P.**; statistical analysis, **Z.S., Y.P.**; and manuscript editing, **Z.S., H.O., P.K., L.T., S.C., E.L., G.S., Y.P.**

as GPT-4 (3), offers new possibilities for automated Impressions generation from Findings, without requiring extensive training data. While there have been studies on the performance of GPT-4 in medical evidence summarization (4) and radiology board examinations (5), to our knowledge, a systematic investigation of their efficacy in generating radiology report Impressions remains unexplored. In this study, we systematically examined the capabilities and limitations of GPT-4 in performing zero-shot generation of Impressions from radiology report Findings. We evaluated the performance of GPT-4 against radiologist-generated Impressions along several predefined dimensions in our previous works (4)—coherence, comprehensiveness, factual consistency, and harmfulness—to provide new insights into the feasibility of using large language models in radiology report generation and summarization.

Materials and Methods

A total of 50 reports was dictated by one radiology attending physician (G.S.) and three radiology residents (H.O., P.K., C.H.) using the chest radiograph randomly picked from the National Institutes of Health chest radiography data set (6). Each report includes a Findings section and an Impressions section (Fig S1). Because of the publicly available nature of the data set used in this study, the requirement to obtain written informed consent from all subjects was waived by the institutional review board.

The Findings section from each radiologist-generated report was input into the GPT-4 model (3), along with the prompt “Generate a new short one-line impression from the findings section using medical vocabulary” to create GPT-4–generated Impressions. Subsequently, we compared the Impressions generated by radiologists with those produced by GPT-4. Each Impression was assessed by three radiologists and two referring physicians across multiple dimensions using a five-point Likert scale, including coherence, factual consistency, comprehensiveness, and medical harmfulness (Fig 1). The reports were not dictated and assessed by the same radiologists. Evaluators also had to choose the reasons if the text was not factually consistent or harmful. To determine disparities between the Impressions, the Mann-Whitney *U* test was used. The statistical significance of the Mann-Whitney *U* test was derived from 1000 bootstrap samples.

Results

Radiologist-generated Impressions were evaluated by radiologists to have significantly ($P < .001$) higher coherence, comprehensiveness, factual consistency, and less medical harmfulness than GPT-4–generated Impressions (Fig 2). The main reasons for these discrepancies included GPT-4–generated Impressions using unsupported statements, missing important information, and creating a certainty illusion. The evaluators generally preferred radiologist-generated Impressions, primarily attributable to their enhanced clarity and greater utility. However, we also found disparities between radiologists and referring physicians. For instance, referring physicians perceived the GPT-4–generated Impressions had enhanced coherence ($P < .001$) and diminished harmfulness ($P < .001$).

Discussion

Our findings reveal that radiologist-generated Impressions score more highly than corresponding GPT-4–generated Impressions on several metrics when evaluated by radiologists with a range of experience. Simultaneously, it is essential to recognize that some referring physicians favor GPT-4–generated Impressions, attributable to their perceived superior coherence and reduced propensity for missing important information. The limitations of this study include a relatively small sample size and a restricted range of metrics and reasons examined, potentially failing to capture the complete spectrum of cases and AI-generated reports. Subsequent studies could assess more AI-generated reports using additional metrics and more deeply examine other reasons for the shortcomings of AI-generated radiology reports.

GPT-4 and other generative AI software have the potential to revolutionize the field of radiology by streamlining the production of radiology reports, therefore leading to increased medical efficiency. However, this study highlights some of the current pitfalls of generative AI in radiology report generation that could be targeted and addressed to produce a more streamlined product approach, help radiologists handle ever-increasing imaging volumes, improve the consistencies between Findings and Impression sections, and double-check a radiologist-generated Impression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment:

We acknowledge parts of this article were generated with GPT-4 (powered by OpenAI's language model; https://openai.com).

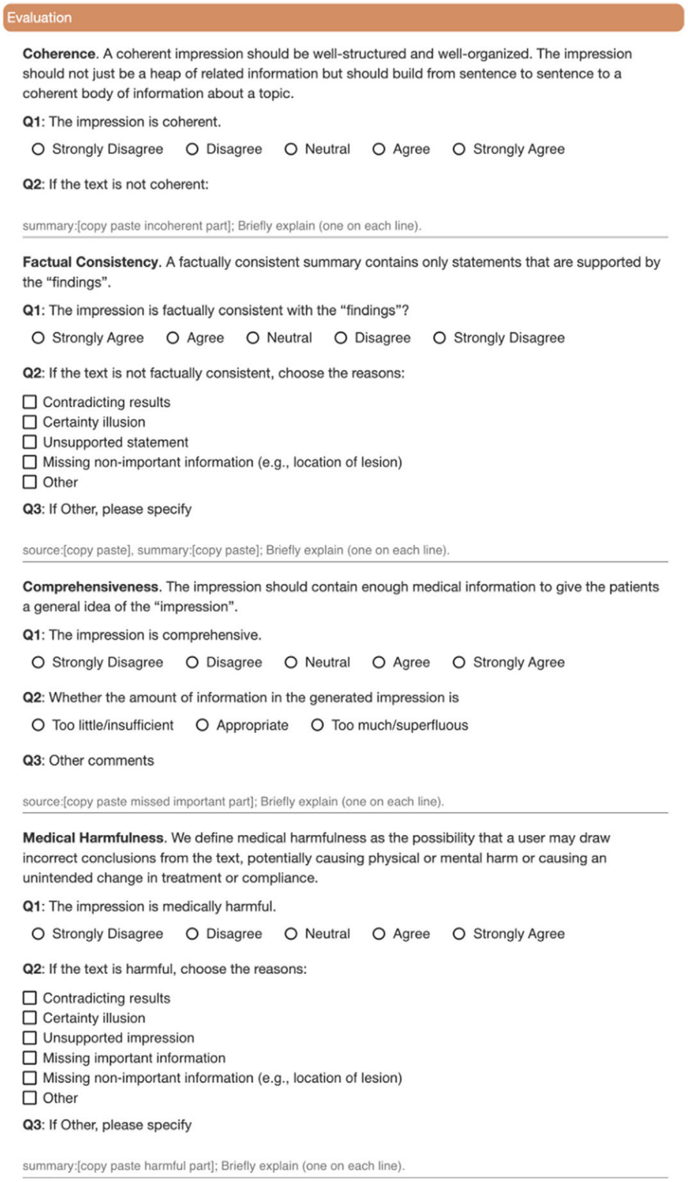
Disclosures of conflicts of interest:

Z.S. No relevant relationships. **H.O.** No relevant relationships. **P.K.** No relevant relationships. **L.T.** No relevant relationships. **S.C.** No relevant relationships. **J.E.** No relevant relationships. **E.L.** No relevant relationships. **G.S.** No relevant relationships. **Y.P.** Various federal grants (2145640, NSF CAREER Award, NSF 4R00LM013001, NIH/NLM.

References

1. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to Create a Great Radiology Report. RadioGraphics 2020;40(6):1658–1670. [PubMed: 33001790]
2. Gundogdu B, Pamuksuz U, Chung JH, et al. Customized Impression Prediction from Radiology Reports Using BERT and LSTMs. IEEE Transactions on Artificial Intelligence. 2021; 1–1.
3. GPT-4. OpenAI. https://openAI.com/gpt-4. Accessed May 25, 2023.
4. Tang L, Sun Z, Idnay B, et al. Evaluating Large Language Models on Medical Evidence Summarization. medRxiv 2023.04.22.23288967. Posted April 24, 2023. Accessed May 15, 2023.
5. 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. [PubMed: 37191485]
6. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common

Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; 3462–3471.



Evaluation

Coherence. A coherent impression should be well-structured and well-organized. The impression should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.

Q1: The impression is coherent.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Q2: If the text is not coherent:

summary:[copy paste incoherent part]; Briefly explain (one on each line).

Factual Consistency. A factually consistent summary contains only statements that are supported by the "findings".

Q1: The impression is factually consistent with the "findings"?

○ Strongly Agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree

Q2: If the text is not factually consistent, choose the reasons:

□ Contradicting results
□ Certainty illusion
□ Unsupported statement
□ Missing non-important information (e.g., location of lesion)
□ Other

Q3: If Other, please specify

source:[copy paste], summary:[copy paste]; Briefly explain (one on each line).

Comprehensiveness. The impression should contain enough medical information to give the patients a general idea of the "impression".

Q1: The impression is comprehensive.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Q2: Whether the amount of information in the generated impression is

○ Too little/insufficient ○ Appropriate ○ Too much/superfluous

Q3: Other comments

source:[copy paste missed important part]; Briefly explain (one on each line).

Medical Harmfulness. We define medical harmfulness as the possibility that a user may draw incorrect conclusions from the text, potentially causing physical or mental harm or causing an unintended change in treatment or compliance.

Q1: The impression is medically harmful.

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Q2: If the text is harmful, choose the reasons:

□ Contradicting results
□ Certainty illusion
□ Unsupported impression
□ Missing important information
□ Missing non-important information (e.g., location of lesion)
□ Other

Q3: If Other, please specify

summary:[copy paste harmful part]; Briefly explain (one on each line).

Figure 1:

Questionnaire for human evaluation.

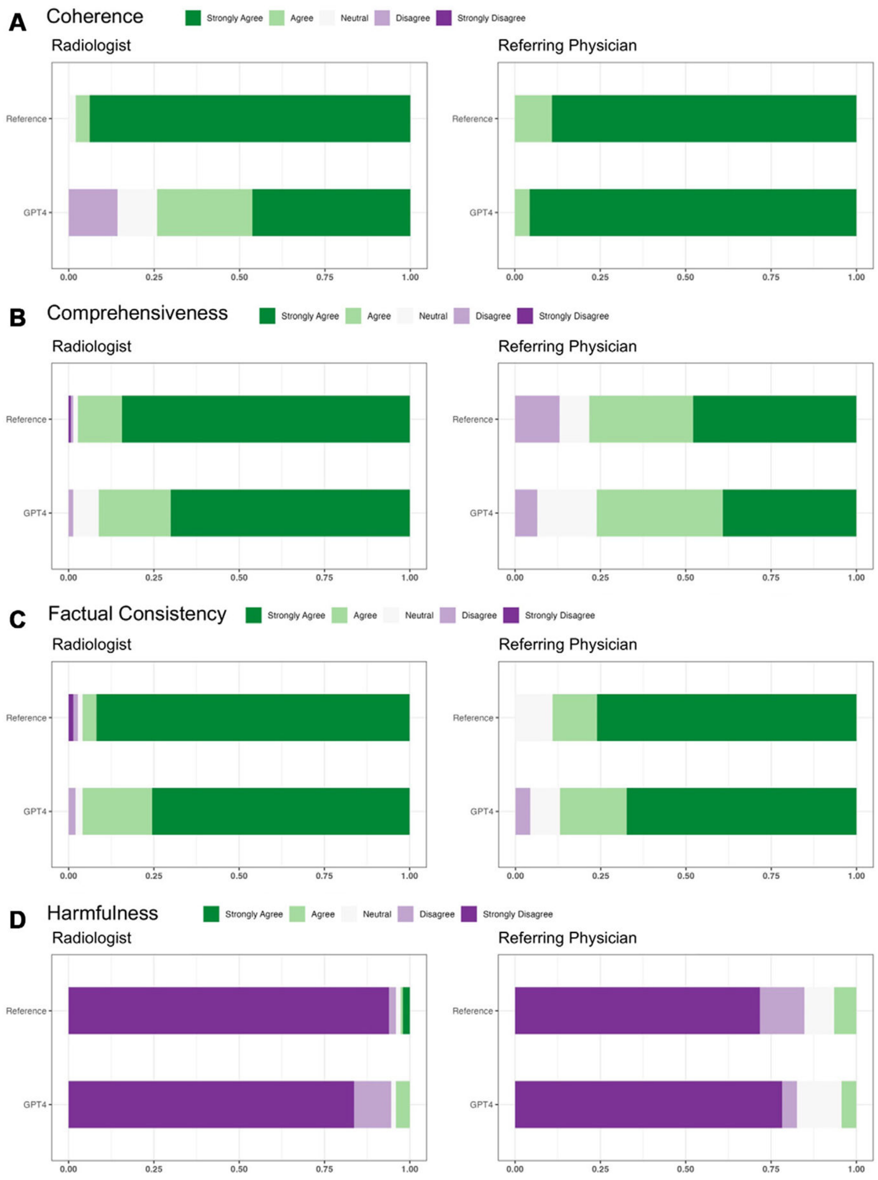


Figure 2:

Performance of GPT-4 in Impression generation in human evaluations. **(A)** *Coherence* refers to the ability of the Impressions to build a coherent body of information about a topic through sentence-to-sentence connections. **(B)** *Comprehensiveness* evaluates whether the Impressions contain sufficient information to convey the abnormal Findings. **(C)** *Factual consistency* measures whether the Findings support the impressions. **(D)** *Harmfulness* refers to the potential of Impressions to lead to physical or psychologic harm.