Research Letter

Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception

Sebastian Ziegelmayer, MD; Alexander W Marka, MD; Nicolas Lenhart, MD; Nadja Nehls, MD; Stefan Reischl, MD; Felix Harder, MD; Andreas Sauter, MD; Marcus Makowski, MD; Markus Graf*, MD; Joshua Gawlitza*, MD

Department of Diagnostic and Interventional Radiology, School of Medicine & Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

*these authors contributed equally

Corresponding Author:
Sebastian Ziegelmayer, MD
Department of Diagnostic and Interventional Radiology
School of Medicine & Klinikum rechts der Isar
Technical University of Munich
Ismaninger Straße 22
Munich, 81675
Germany
Phone: 49 1759153694
Email: ga89rog@mytum.de

Abstract

Exploring the generative capabilities of the multimodal GPT-4, our study uncovered significant differences between radiological assessments and automatic evaluation metrics for chest x-ray impression generation and revealed radiological bias.

(J Med Internet Res 2023;25:e50865) doi: 10.2196/50865

KEYWORDS

generative model; GPT; medical imaging; artificial intelligence; imaging; radiology; radiological; radiography; diagnostic; chest; x-ray; x-rays; generative; multimodal; impression; impressions; image; images; AI

Introduction

Generative models trained on large-scale data sets have demonstrated an unprecedented ability to generate humanlike text [1] and have performed surprisingly well on untrained tasks (zero-shot learning) [2]. In medical imaging, the applications are manifold, and it has been shown that models can not only draw radiological conclusions [3] but also structure reports [4] and even generate impressions based on the findings given in a report [5] or the image itself [6]. One of the leading obstacles limiting the development of models for generating clinically applicable reports is the lack of evaluation metrics that capture the core aspects of radiological impressions [7,8]. While there are initial studies on the perception of artificial intelligence (AI)–generated text in the general population [9], insights are missing for specialized areas such as medical imaging. Therefore, our study investigated the ability of GPT-4 to generate radiological impressions based on different inputs, focusing on the correlation between radiological assessment of impression quality and common automated evaluation metrics, as well as radiological perception of AI-generated text.

Methods

Overview

To generate and evaluate impressions of chest x-rays based on different input modalities (image, text, text and image), a blinded radiological report was written for 25 cases from a publicly available National Institutes of Health data set [10]. The GPT-4 model was given an image, the results, or both sequentially to generate an input-dependent impression. In a blind randomized reading, 4 radiologists rated the impressions based on "coherence," "factual consistency," "comprehensiveness," and "medical harmfulness," which were used to generate a radiological score based on a 5-point Likert scale of each dimension. Additionally, radiologists were asked to classify the origin of the impression (human, AI), providing justification for their decision. The text model evaluation metrics and their correlation with the radiological score were assessed. Lastly, common model metrics for text evaluation were extracted and compared to the radiological assessment. The supplementary methods in Multimedia Appendix 1 [5,8,10-17] provide further details.

Ethical Considerations

Due to the publicly available data set used in this study, the requirement to obtain written informed consent from the participants was waived. Participants were anonymized.

Results

According to the radiological score, the human-written impression was rated highest, although not significantly higher than the text-based impressions (Table 1). A detailed analysis is shown in the supplementary results section in Multimedia Appendix 1. The automated evaluation metrics showed moderate correlations to the radiological score for the image impressions; however, individual scores diverged depending on the input (Figure 1). Correct detection of an impression's origin (human/AI) varied by input (text: 61/100, 61%; image: 87/100, 87%; radiologist: 87/100, 87%; text and image: 63/100, 63%). For the text input, a homogeneous distribution was found, similar to radiological impressions classified as AI generated (supplementary figure in Multimedia Appendix 1). It was shown that impressions classified as human written were rated significantly higher by the radiologist, with a mean score of 18.11 (SD 1.87) for impressions classified as human written and 13.41 (SD 3.93; $P \leq .001$) for impressions classified as AI generated.

Table 1. Quantitative and qualitative scores based on the input[a].

	Qualitative			Quantitative		
	Radiologist score	BLEU[b]	BERT[c]	CheXbert vector similarity	RadGraph	RadCliQ
Image	10.97[d]	0.051[e]	0.298[e]	0.471	0.038[d]	0.328[d]
Text	16.95	0.125	0.356	0.417	0.168	0.291
Text and image	15.54[d]	0.173	0.411	0.523	0.197	0.278
Radiologist	18.47	N/A[f]	N/A	N/A	N/A	N/A

[a]Except for RadCiQ, which corresponds to the error rate, a higher score indicates a better approximation. For the automated metrics, the text and image–based impression score was highest, while the radiological score for the text-based impression was closest to the radiological ground truth.

[b]BLEU: bilingual evaluation understudy.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]Indicates a *P* value <.05 for all higher input scores.

[e]Indicates a *P* value <.05 compared to the highest score.

[f]N/A: not applicable.

Figure 1. Scatterplots for each automated metric (BERT=blue; BLEU=yellow; CheXbert vector similarity=gray; RadGraph=light blue; RadCliQ=red) depending on the input: (A) image, (B) text, or (C) text and image. For the image input, all metrics except CheXbert vector similarity showed a significant correlation. However, the correlation was divergent or opposing for the text and text and image inputs. All correlation coefficients with their *P* values are shown in the lower section of the figure. BERT: Bidirectional Encoder Representations From Transformers; BLEU: bilingual evaluation understudy.

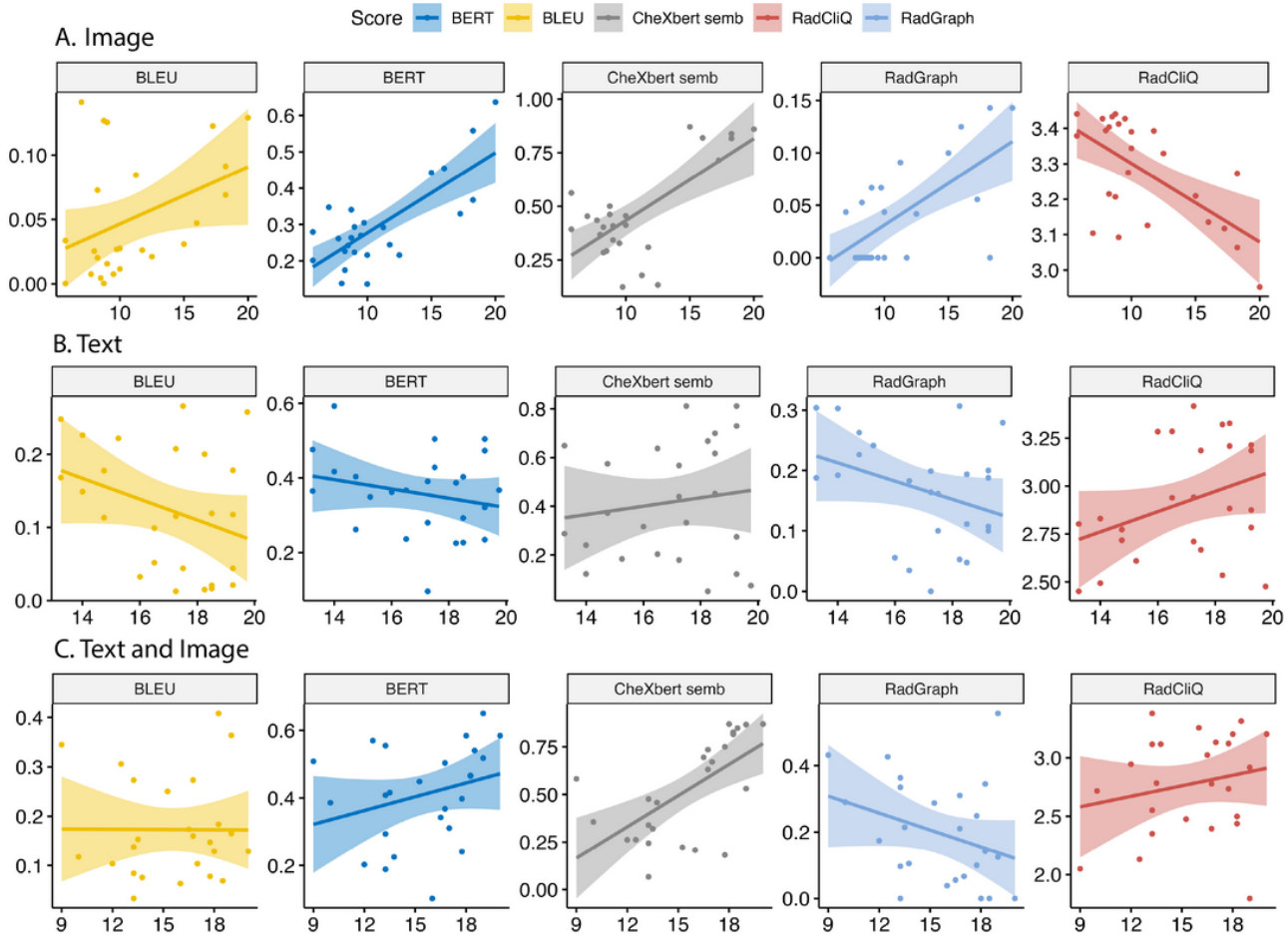

	Image	*P*	*Text*	*P*	*Text and Image*	*P*
BLEU	0.29	.04	-0.18	.21	0.027	.85
BERT	0.37	.01	-0.093	.53	0.3	.04
CheXbert vector similarity	0.19	.02	0.1	.50	0.49	<.001
RadGraph	0.47	.003	-0.18	.21	-0.25	.09
RadCliQ	-0.41	.004	0.18	.21	0.17	.25

Discussion

We evaluated the "out-of-the-box" performance of GPT-4 for chest x-ray impression generation based on different inputs. Based on the radiological score, text-based impressions were not significantly lower than the radiological impressions, whereas other inputs were rated significantly lower. Sun et al [5] showed that text-based impressions rated by radiologists were inferior. However, the study did not clarify if the radiological evaluations of the impressions were conducted under blinded conditions. Our work identified radiological bias, as impressions classified as human written received higher ratings. Therefore, without blinding, there is a risk that the inferiority of the AI-generated impressions is due to bias.

For the automated metrics, the impressions based on text and image were rated the closest to the radiological impressions, followed by text-based impressions. For the image-based impressions, there was a significant moderate correlation between the automated metrics and the radiological score; however, for the other inputs, opposite or nonsignificant

correlations were found. Automatic metrics that capture relevant aspects of report quality are a prerequisite for successful development and clinical integration. Evaluation metrics, however, can only be as good as the human assessment, which is not free of bias and characterized by false heuristics [9]. Our findings underline this point, as impressions that were classified as human written scored significantly higher in the radiological assessment. Human evaluation is not error-free, but it is the benchmark for the evaluation of generated text.

Radiological heuristics, sources of error, and relevant aspects of radiological quality need to be further investigated, as they are essential for the development of useful model metrics.

Acknowledgments

No generative model was used to write, edit, or review the manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed methods, detailed results, and a mosaic plot visualizing the justification for classifying an impression as artificial intelligence generated.
[DOCX File , 5377 KB-Multimedia Appendix 1]

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [FREE Full text]
2. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. 2021 Presented at: 38th International Conference on Machine Learning; July 18-24, 2021; Virtual
3. Bhayana R, Bleakney R, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology. 2023 Jun;307(5):e230987 [doi: 10.1148/radiol.230987] [Medline: 37191491]
4. Adams L, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology. 2023 May;307(4):e230725 [doi: 10.1148/radiol.230725] [Medline: 37014240]
5. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology. 2023 Jun;307(5):e231259 [FREE Full text] [doi: 10.1148/radiol.231259] [Medline: 37367439]
6. Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. 2021 Presented at: Machine Learning for Health; December 4, 2021; Virtual
7. Hartung M, Bickle I, Gaillard F, Kanne JP. How to create a great radiology report. Radiographics. 2020 Oct;40(6):1658-1670 [doi: 10.1148/rg.2020200020] [Medline: 33001790]
8. Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns (N Y). 2023 Sep 08;4(9):100802 [FREE Full text] [doi: 10.1016/j.patter.2023.100802] [Medline: 37720336]
9. Jakesch M, Hancock JT, Naaman M. Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci U S A. 2023 Mar 14;120(11):e2208839120 [FREE Full text] [doi: 10.1073/pnas.2208839120] [Medline: 36881628]
10. 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. Proc IEEE Conference Computer Vis Pattern Recognition. 2017:2097-2106 [doi: 10.1109/cvpr.2017.369]
11. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. Preprint posted online on May 22, 2020. 2023
12. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on January 28, 2022. 2023
13. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv. Preprint posted online on May 24, 2022. 2023
14. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. 2002 Presented at: 40th Annual Meeting of the Association for Computational Linguistics; July 2002; Philadelphia, PA
15. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv. Preprint posted online on April 21, 2019. 2023

16. Smit A, Jain S, Rajpurkar P, Pareek A, Ng A, Lungren M. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. 2020 Presented at: Conference on Empirical Methods in Natural Language Processing; November 2020; Online
17. Jain S, Agrawal A, Saporta A, Truong SQH, Duong DN, Bui T, et al. RadGraph: extracting clinical entities and relations from radiology reports. arXiv. Preprint posted online on June 28, 2021. 2023

Abbreviations

AI: artificial intelligence

Edited by A Mavragani; submitted 14.07.23; peer-reviewed by R Toomey; comments to author 15.08.23; revised version received 16.08.23; accepted 27.11.23; published 22.12.23

Please cite as:
Ziegelmayer S, Marka AW, Lenhart N, Nehls N, Reischl S, Harder F, Sauter A, Makowski M, Graf M, Gawlitza J
Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception
J Med Internet Res 2023;25:e50865
URL: https://www.jmir.org/2023/1/e50865
doi: 10.2196/50865
PMID: 38133918