This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

MRScore: Evaluating Medical Report with LLM-based Reward System

Yunyi Liu[1], Zhanyu Wang[1], Yingshu Li[1], Xinyu Liang[2], Lingqiao Liu[3], Lei Wang[4], and Luping Zhou[1](✉)

[1] The University of Sydney, Sydney, NSW, Australia
{yunyi.liu1,zhanyu.wang,yingshu li, luping.zhou}@sydney.edu.au
[2] Guangzhou University of Chinese Medicine, Guangzhou, China
xinyu.liang31@gmail.com
[3] The University of Adelaide, Adelaide, SA, Australia
lingqiao.liu@adelaide.edu.au
[4] The University of Wollongong, Wollongong, NSW, Australia
leiw@uow.edu.au

Abstract. We propose MRScore, an innovative automatic evaluation metric specifically tailored for the generation of radiology reports. Traditional (natural language generation) NLG metrics like BLEU are inadequate for accurately assessing reports, particularly those generated by Large Language Models (LLMs). Our experimental findings give systematic evidence of these inadequacies within this paper. To overcome this challenge, we have developed a unique framework intended to guide LLMs in evaluating radiology reports, which was created in collaboration with radiologists adhering to standard human report evaluation procedures. Using this as a prompt can ensure that the LLMs' output closely mirrors human analysis. We then used the data generated by LLMs to establish a human-labeled dataset by pairing them with accept and reject samples, subsequently training the MRScore model as the reward model with this dataset. MRScore has demonstrated a higher correlation with human judgments and superior performance in model selection when compared with traditional metrics. Our code is available on GitHub at: https://github.com/yunyiliu/MRScore.

Keywords: Radiology Report Generation · Evaluation metrics · Large Language Models · Reward Model.

1 Introduction

Automated assessment of text generation systems, such as those used in radiology report generation, typically involves the comparison of generated reports against reference reports to gauge semantic accuracy. However, widely utilized metrics, such as the BLEU metric [16], primarily quantify n-gram matches, thereby neglecting the critical aspects of lexical and structural diversity that are essential for preserving meaning. Currently, there are two typical shortcomings found in n-gram-based evaluation metrics [10]. Firstly, these metrics often misjudge paraphrasing due to rigid pattern matching. For instance, traditional metrics, such as the BLEU and METEOR [18], may erroneously favor a radiology report's expression like "patient exhibits no symptoms" over a semantically identical phrase "symptom-free patient." This discrepancy arises because these metrics penalize deviations from the reference structure, regardless of semantic equivalence. Various approaches address this challenge. For instance, Bert Score [25], in contrast, calculates similarity with contextualized token embedding, proven to detect paraphrasing more effectively. Secondly, traditional n-gram approaches can miss critical semantic nuances in sentence structure. For example, if one report states "No evidence of pathology was observed following the MRI scan," and another says "Following the MRI scan, no evidence of pathology was observed," BLEU will inadequately penalize this variation, despite both sentences conveying the same meaning. Contextualized embeddings, however, are adept at capturing such nuances in sentence structure and order.

This study introduces MRScore, an innovative metric designed for evaluating automated radiology report generation. Developed in collaboration with professional radiologists, MRScore is underpinned by a framework that articulates their expert rules and priorities for report assessment. Our analysis first identified the limitations of existing evaluation metrics. To address these gaps, we created MRScore as a bespoke framework for evaluating radiology reports. We trained MRScore using a reward model, which necessitated the development of a human-ranked dataset. This was achieved by employing our error-based evaluation framework as a prompt to guide GPT-4 in generating human-like evaluations. Using this framework and 1,000 ground truth reports, we generated 3,000 predicted datasets across three distinct scoring levels in seven criteria outlined in our evaluation framework. These will be detailed later in this paper. We assessed the human correlation of this dataset by having a radiologist score 100 randomly selected reports. With confirmed high human correlation, this dataset was then used to train our reward model. During the training preparation, we paired the reports as accept, reject, with 'accept' denoting the report with the higher score and 'reject' the lower. We also introduced a margin to indicate the score difference between the paired reports, providing the model with a measure of the distance between accepted and rejected reports. For training, we employed Mistral-7B-instruct [8] as our pre-trained model. To validate our model, we scored 100 sample reports generated by GPT-4V and compared these scores with those from other existing evaluation methods. Our correlation calculations showed that MRScore achieved a higher alignment with human judgment than other metrics.

Our main contributions are summarized as follows:

(1) The paper critically evaluates traditional NLG metrics(e.g., BLEU [16], CIDEr [19])for LLM-generated text, noting their inconsistency with human evaluations.
(2) A novel, error-based framework is introduced, transforming radiologists' evaluation criteria into a binary, weighted scoring system across seven stan-

dards, significantly aligning model outputs with human assessments, evidenced by Kendall's tau of 0.65.

(3) Leveraging this framework, we trained MRScore, an LLM-based model for automated report scoring, which outperformed other metrics in human correlation, achieving Kendall's tau of 0.250 and Spearman's coefficient of 0.304.

2 Problem Statement and Prior Metrics

In the domain of automated radiology report generation, the efficacy of generated reports is assessed through a metric function $f(x, \hat{x})$, where x represents the generated report and $\hat{x}$ is the reference report. Traditional evaluation metrics such as BLEU [16], ROUGE [12], METEOR [1], and CIDEr [19] are commonly employed for this purpose. However, these metrics predominantly rely on n-gram overlap, which may not adequately capture semantic equivalence between the generated and reference texts. Our study meets the need for a more sophisticated metric that can evaluate the semantic content and clinical relevance of radiology reports more accurately.

To systematically demonstrate the limitations of traditional metrics in evaluating radiology report quality, we utilized GPT-4V to generate reports for the entire MIMIC-CXR dataset, subsequently computing the NLG scores (e.g., BLEU, ROUGE, METEOR, and CIDEr) for these reports. From this comprehensive dataset, we meticulously selected 100 reports to undergo detailed human evaluation. We then calculated the correlation between these human evaluations and the traditional NLG scores, aiming to highlight the discrepancies and underline the inadequacy of traditional metrics in capturing the nuances of clinical reporting.

Traditional NLG Metrics We evaluated GPT-4V-generated results using traditional metrics, comparing them with state-of-the-art (SOTA) benchmarks. Table 1 details this performance comparison on the MIMIC-CXR dataset [9], focusing on radiology report generation methods. It shows that GPT-4V scores are very low on all conventional metrics. To verify the efficiency of these scores, we conducted the human evaluation in the following section.

Table 1. Comparison on the MIMIC-CXR dataset.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
R2Gen [3]	0.353	0.218	0.145	0.103	0.277	0.142	-
R2GenCMN [2]	0.353	0.218	0.148	0.106	0.278	0.142	-
PPKED [13]	0.360	0.224	0.149	0.106	0.284	0.149	0.237
GSK [23]	0.363	0.228	0.156	0.115	0.284	-	0.203
MSAT [21]	0.373	0.235	0.162	0.120	0.282	0.143	0.299
METransformer [20]	**0.386**	**0.250**	**0.169**	**0.124**	**0.291**	**0.152**	**0.362**
GPT-4V [26]	0.338	0.190	0.109	0.061	0.240	0.125	0.033

Human Correlation Analysis We analyzed 100 report pairs, comprising ground truth and GPT-4V-generated reports, graded by a radiologist into high(90), medium(60), and low(30) tiers. We compared these human ratings with NLG metrics (e.g., BLEU, ROUGE, METEOR, CIDEr) and assessed their correlation using Kendall's τ coefficient and Spearman's ρ coefficient, represented as:

$$\begin{gathered}\tau=\frac{\text{number of concordant pairs}-\text{number of discordant pairs}}{\text{total number of pairs}\times(\text{total number of pairs}-1)/2},\\ \rho=1-\frac{6\sum d_i^2}{n(n^2-1)},\end{gathered}\tag{1}$$

where d_i is the difference between the ranks of corresponding variables and n is the number of observations.

Table 3 demonstrates near-zero correlation coefficients and high p-values when compared to human evaluations, suggesting a minimal correlation with human judgment. Given that effective metrics should highly correlate with human evaluations, these results imply their unsuitability. For illustrative purposes, Fig 1 highlights an example where a report received a BLEU score of 0.069e-6, indicative of a low NLG evaluation score, yet was highly rated by a professional radiologist.

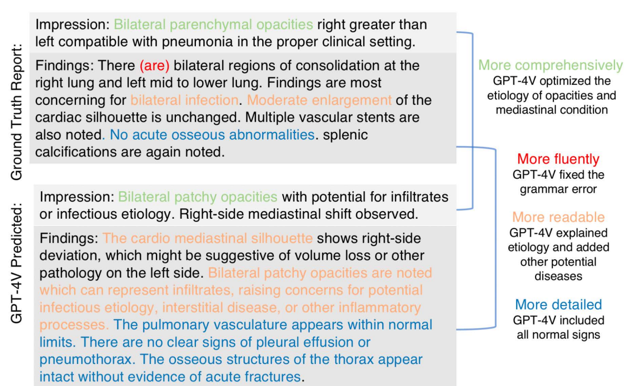

Fig. 1. An example of a ground truth report and a GPT-4V generated report. Key medical information in the reports is highlighted using different colors.

Building on the preceding analysis, we introduce MRScore, an innovative evaluation metric refined through a reward model within our novel evaluation framework. The methodology is elaborated in subsequent sections.

3 Method

In this work, a well-trained radiologist helped develop a scoring system based on seven key rules derived from expertise and academic research, ensuring its reliability. Integrating this system with GPT-4 for report evaluation, we achieved outputs highly correlated with human judgments. This validated framework enabled us to create a dataset reflecting human evaluative preferences, crucial for fine-tuning our reward model, which has proven to outperform conventional NLG metrics in aligning with human assessment standards.

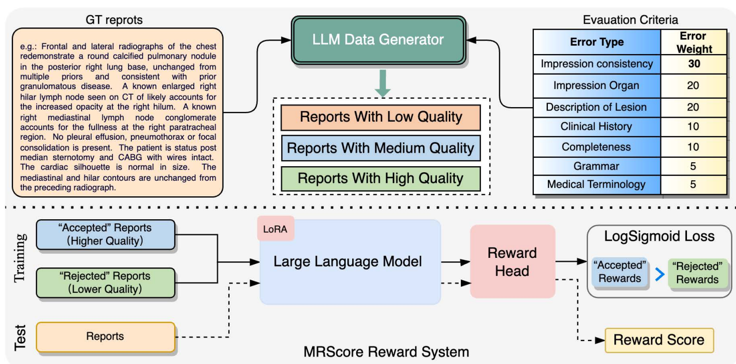

Fig. 2. Overview of MRScore: The upper part is the process of generating data, the lower part is the process of training the reward model by using LoRA, the dashed line is the testing phase, and the solid line is the training phase.

3.1 Report Evaluation Criteria Definition

We introduced a new error-based evaluation framework, validated through radiologist assessments and literature, ensuring robust report evaluation. The method assesses criteria sequentially, adjusting weights for detected errors, with specifics outlined in Table 2. A comprehensive analysis of each error category and design specifics follows.

3.2 Reward Model

This part will briefly introduce our training process.

Start with a pretrained Model Mistral-7B-instruct is utilized as LLM. This model provides a solid foundation for language understanding and strong capabilities in language comprehension.

Generate Training Data In this phase the pertained model is prompted with prompts x to produce pairs of answers $(y_1, y_2) \sim \pi^{SFT}(y|x)$. These are then presented to human labelers who express preferences for one answer, denoted as $y_w > y_l|x$ where y_w and y_l denotes the preferred and dispreferred completion amongst (y_1, y_2) respectively. Here we used GPT-4V to replace the human labeler, we generated reports with scores using our innovative error-based evaluation framework, simulating how a human ranks these reports.

Table 2. Table for error types and design detail.

Error Type and weight	Design Detail
Impression consistency:30	Assess 'impression' section's presence for crucial diagnostic details, vital for quality care[6].
Impression Organ:20	Evaluate impression precision and detail on affected organs, as per standards[4].
Description of Lesion:20	Ensure accurate lesion description, including location, size, and related details, reflecting ground truth accuracy.
Clinical History:10	Confirm the report reflects accurate clinical history, integrating patient history with imaging findings[17].
Completeness:10	Check report completeness, a critical aspect reflecting radiologists' expertise.
Grammar:5	Guarantee report's grammatical accuracy, ensuring clarity and preventing misinterpretation[22,15].
Medical Terminology:5	Ensure proper use of medical terminology, key for clear healthcare communication[14].

Define the Objective The model is trained to predict human preferences accurately. This usually involves defining a reward function that the model aims to maximize. For MRScore, the reward function is derived from the rankings of radiology reports, with the model learning to predict the more preferred report in each pair.

Fine-Tuning This fine-tuning adapts the model to the specifics of the reward prediction, aligning its outputs with the expected human evaluations. In our MRScore training, we fine-tuned the Mistral-7B-instruct model on our paired dataset, teaching it to distinguish between higher and lower-quality reports based on the derived scores. $margin = score_{accpet} - score_{reject}$. The reward head is a linear projecting layer that will project the feature to 1 dimension, there will be a Sigmoid function to get the final reward. There is a simple process graph in 3.

Evaluation Finally, assess the trained model's performance to ensure it aligns with human judgment or the desired outcomes. For MRScore, we evaluated the model's effectiveness by its alignment with expert radiologist evaluations, ensuring the model's predictions correlate strongly with human expert rankings.

3.3 Loss Function

The equation 2 is the loss function for our reward model. The γ_θ is the reward return back by the reward model. y_w represents higher value reports, and the y_l represents the lower value reports. The log is the logistic function. D is the reference dataset. m represents the margin. K represents the batch size.

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} \sum_{(x,y_w,y_l)\sim D} \left[\log\left(\sigma\left(\gamma_\theta(x,y_w) - \gamma_\theta(x,y_l) - m\right)\right)\right] \tag{2}$$

4 Experiments and Result

4.1 Dataset

Scoring Data Utilizing the new radiology report generation framework, we employed the GPT4 API to create 3000 datasets with varied scores, stratified into three tiers to maintain balanced data distribution: 0-40, 40-70, and 70-100, ensuring uniform coverage across the scoring spectrum.

Paired Data We generated paired data with human rankings from the scoring dataset, assigning higher-scored reports as accepted and lower-scored as rejected, using the score difference as the margin. This method ensures the dataset reflects human ratings, aiding the model in learning rewards and penalties. The dataset comprises 2598 training and 200 testing entries.

Evaluate Human Correlation for the Scoring Data We analyzed 100 GPT-generated samples with radiologist evaluations to measure the human correlation. The Pearson correlation of 0.65 indicates significant agreement between radiologist and GPT scores, affirming the reliability of GPT's scoring data for reward model training. More details will be provided in the supplementary.

4.2 Experiment Result

In our study's second phase, we assessed the human correlation for 100 samples against traditional metrics (e.g., Bleu, Rouge, Meteor, Cider) and observed low correlations, suggesting their limited evaluation effectiveness. Subsequently, we examined the correlation of our MRScore with more metrics like BertScore, and RadgraphF1, known for their semantic evaluation efficacy. Our findings, detailed in Table 3, indicate MRScore's superior correlation with human judgments, evidenced by Kendall's Tau (0.250) and Spearman's coefficient (0.304), outperforming traditional NLG metrics and showcasing the strongest alignment. While traditional metrics showed insignificant correlations, Bert Score [24], Radgraph F1 [7], and MRScore presented statistically significant correlations with lower P-values. Our MRScore has the best preference with the highest correlation.

Table 3. Evaluation of P-Value, Kendall's Tau and Spearman coefficient

	Bleu-4	ROUGE_L	METEOR	CIDEr	Bert Score [24]	Radgraph F1 [7]	**MRScore**
P Value ↓	0.688	0.429	0.460	0.503	0.0446	0.071	**0.002**
Kendall's Tau ↑	0.032	0.063	0.059	0.053	0.159	0.144	**0.250**
P Value ↓	0.677	0.484	0.463	0.422	0.045	0.08	**0.002**
Spearman ↑	0.042	0.071	0.074	0.081	0.200	0.176	**0.304**

The scatter plots in Fig. 3 present comparisons between various scoring metrics and human evaluation rates for radiology reports. Each plot shows a different metric, with points indicating individual report scores against human ratings.

The trends are represented by lines with shaded areas demonstrating confidence intervals. MRScore appears to have the most positive correlation with human ratings, suggesting that it closely aligns with professional evaluations in the medical report. Traditional NLG metrics like Bleu-4, METEOR, and ROUGE_L show some positive correlation with human ratings but with a greater spread, indicating variability in their alignment. The Bertscore also seems to positively correlate with human ratings, though to a lesser degree than MRScore. Radgraph F1 shows a positive trend but not as strong as MRScore. Overall, MRScore stands out as a promising metric for aligning with human judgment, potentially indicating its effectiveness.

Table 4 presents the results of different LLMs as base models trained on our preference dataset and reward models, along with their correlation with human scores. Two correlation scores are used: Kendall and Spearman correlation coefficients. The Mistral-7b performs the best in terms of consistency with the human ratings model and has 6.8M trainable parameters, with a Kendall correlation of 0.179 and a Spearman correlation of 0.220. So, we selected Mistral as our based LLM.

Table 4. Human Evaluation Result of Different LLM base models

Model	Trainable params (%)	Kendall ↑(P value ↓)	Spearman ↑(P value ↓)
Phi-1.5 [11]	5.2M (0.197)	0.153 (0.056)	0.192 (0.055)
Gemma-2b-it [5]	1.8M (0.073)	0.135 (0.091)	0.169 (0.092)
Gemma-7b-it [5]	6.4M (0.075)	0.170 (0.034)	0.209 (0.037)
Mistral-7b [8]	6.8M (0.096)	**0.250** (**0.002**)	**0.304** (**0.002**)

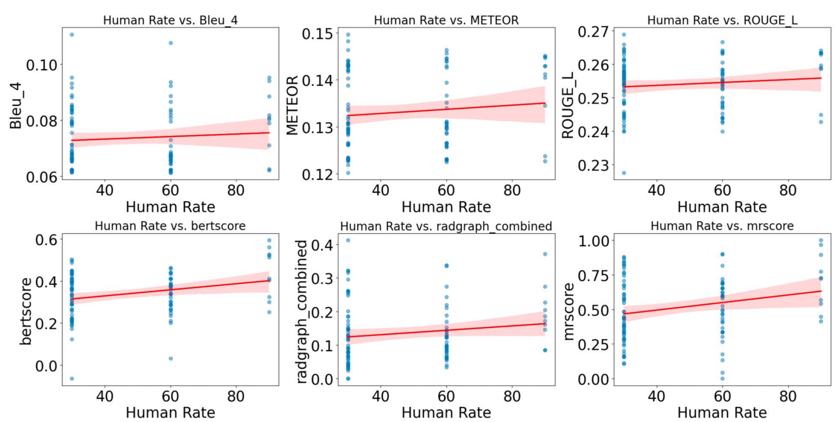

Fig. 3. Correlation between metrics score and radiologist score

5 Conclusion

In conclusion, MRScore stands as an innovative metric that significantly enhances the evaluation of radiology reports generated by LLMs, aligning closely with human expert evaluations. Its design, rooted in an error-based evaluation framework co-developed with radiologists, ensures a strong correlation with human judgment. Demonstrating higher correlation coefficients (Kendall's tau of 0.250 and Spearman's 0.304) than traditional metrics.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
4. Ganeshan, D., Duong, P.A.T., Probyn, L., Lenchik, L., McArthur, T.A., Retrouvey, M., Ghobadi, E.H., Desouches, S.L., Pastel, D., Francis, I.R.: Structured reporting in radiology. Academic radiology **25**(1), 66–73 (2018)
5. Gemma Team, T.M., Hardin, C., Dadashi, R., Bhupatiraju, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P., Hussenot, L., et al.: Gemma (2024). https://doi.org/10.34740/KAGGLE/M/3301, https://www.kaggle.com/m/3301
6. Hartung, M.P., Bickle, I.C., Gaillard, F., Kanne, J.P.: How to create a great radiology report. RadioGraphics **40**(6), 1658–1670 (2020)
7. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., et al.: Mistral 7b (2023)
9. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
10. Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., Zhou, L.: A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. medRxiv pp. 2023–11 (2023)
11. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are all you need ii: phi-1.5 technical report (2023)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
13. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
14. Lukaszewicz, A., Uricchio, J., Gerasymchuk, G.: The art of the radiology report: practical and stylistic guidelines for perfecting the conveyance of imaging findings. Canadian Association of Radiologists Journal **67**(4), 318–321 (2016)
15. Pahadia, M., Khurana, S., Geha, H., Deahl, S.T.I.: Radiology report writing skills: A linguistic and technical guide for early-career oral and maxillofacial radiologists. Imaging Science in Dentistry **50**(3), 269 (2020)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Radiology, C.C.: Why clinical history is essential for diagnoses. https://radiologyblog.cincinnatichildrens.org/why-clinical-history-essential-for-diagnoses/ (2017), accessed: 2023-02-24

18. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410, https://doi.org/10.18653/v1/D19-1410
19. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
20. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023)
21. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
22. Wilcox, J.R.: The written radiology report. Applied Radiology **35**(7) (2006)
23. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Radiology report generation with general and specific knowledge. Medical Image Analysis (2021)
24. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
25. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr
26. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023)