Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065992

ARTICLE

A Semantic Evaluation Framework for Medical Report Generation Using Large Language Models

Haider Ali, Rashadul Islam Sumon, Abdul Rehman Khalid, Kounen Fathima and Hee Cheol Kim*

Department of Digital Anti-Aging Healthcare, Inje University, Gimhae, 50813, Republic of Korea

*Corresponding Author: Hee Cheol Kim. Email: heeki@inje.ac.kr

Received: 27 March 2025; Accepted: 11 June 2025; Published: 30 July 2025

ABSTRACT: Artificial intelligence is reshaping radiology by enabling automated report generation, yet evaluating the clinical accuracy and relevance of these reports is a challenging task, as traditional natural language generation metrics like BLEU and ROUGE prioritize lexical overlap over clinical relevance. To address this gap, we propose a novel semantic assessment framework for evaluating the accuracy of artificial intelligence-generated radiology reports against ground truth references. We trained 5229 image–report pairs from the Indiana University chest X-ray dataset on the R2GenRL model and generated a benchmark dataset on test data from the Indiana University chest X-ray and MIMIC-CXR datasets. These datasets were selected for their public availability, large scale, and comprehensive coverage of diverse clinical cases in chest radiography, enabling robust evaluation and comparison with prior work. Results demonstrate that the Mistral model, particularly with task-oriented prompting, achieves superior performance (up to 91.9% accuracy), surpassing other models and closely aligning with established metrics like BERTScore-F1 (88.1%) and CLIP-Score (88.7%). Statistical analyses, including paired t-tests ($p < 0.01$) and analysis of variance ($p < 0.05$), confirm significant improvements driven by structured prompting. Failure case analysis reveals limitations, such as over-reliance on lexical similarity, underscoring the need for domain-specific fine-tuning. This framework advances the evaluation of artificial intelligence-driven (AI-driven) radiology report generation, offering a robust, clinically relevant metric for assessing semantic accuracy and paving the way for more reliable automated systems in medical imaging.

KEYWORDS: Semantic assessment; AI-generated radiology reports; large language models; prompt engineering; semantic score evaluation

1 Introduction

The integration of artificial intelligence (AI) in healthcare has transformed and strengthened clinical data analysis and interpretation. Among its many applications, automating radiology report generation has emerged as a promising approach to streamline workflow and reduce the burden on radiologists [1,2]. In particular, generating precise and semantically accurate reports from medical images such as chest X-rays is essential for supporting timely and informed clinical decision-making [1,3]. Recent advances have emphasized not only the syntactic generation of observations but also the importance of modeling fine-grained attributes (e.g., small pleural effusion) and temporal progression to ensure clinically meaningful outputs [1]. Comprehensive surveys have further highlighted the development of dataset benchmarks, diverse deep learning approaches, and multimodal fusion techniques aimed at improving the performance and reliability of automatic radiology report generation systems [3]. Despite advances in natural language

generation (NLG) and Clinical Efficacy (CE) metrics, assessing the semantic accuracy of these reports remains difficult, especially in high-stakes fields such as medicine.

Radiology reports require a high level of accuracy since they include critical diagnostic information such as lung volumes, pleural effusion, heart size, and lung markings. Any deviation or absence in the generated reports can result in incorrect diagnosis or inadequate treatment options [4]. Diagnostic imaging is an essential and indispensable part of medical diagnosis and treatment, and diagnostic errors or biases are also common in the department of radiology, sometimes even having a severe impact on the diagnosis and treatment of patients [5]. To counter this, it is critical to have a strong evaluation mechanism that goes beyond the traditional NLG metrics such as BLEU [6], METEOR [7], ROUGH [8], and CIDEr [9], which do not consider the semantic level of similarity and only focus on the surface lexical similarity. Semantic alignment with ground truth reports must be emphasized to effectively assess the clinical relevance of generated reports.

Recent developments, specifically in the form of large language models (LLMs), give a good solution to counter the problem of measuring the accuracy of AI-generated radiology reports. By leveraging the true potential of LLMs such as Llama [10], Mistral 7B [11], Gemma 2 [12], and Phi [13], etc., the gap between linguistic fluency and domain-specific clinical relevance between the ground truth radiological and generated radiological reports can be minimized. Moreover, LLMs have shown significant promise in disease and cancer detection, enhancing diagnostic accuracy and clinical decision-making. For instance, specialized LLMs like CancerLLM [14] have been developed to extract cancer phenotypes, generate diagnoses, and propose treatment plans, leveraging pre-training on extensive clinical notes and pathology reports across multiple cancer types. These models improve the precision of oncology workflows by providing efficient and contextually relevant insights [15]. In addition to this, different prompt engineering techniques [16] can also be employed to enhance the accuracy of the generated reports.

In this study, we address the critical challenge of evaluating the semantic accuracy of AI-generated radiology reports by introducing a novel evaluation framework that leverages LLMs and advanced prompt engineering techniques. This proposed framework has the potential to use as an evaluation metric while testing the models in the domain of medical report generation, as traditional metrics mostly don't consider semantic meaning. Using the IU-Xray dataset and the R2GenRL [17] model, we trained a model and a benchmark dataset using MIMIC-CXR and test data of IU-Xray, and employ LLMs (Llama 3.2, Mistral 7B, Phi 3 Medium, and Gemma 2) to assess the semantic alignment of generated reports with ground truth reports. Our approach integrates three distinct prompt engineering strategies, i.e., Zero shot, Chain of Thought [18], and Tree of Thoughts [19], to enhance the precision of semantic scoring. This work advances the state-of-the-art by proposing a semantic scoring metric (0 to 10 scale) that bridges the gap between traditional syntactic metrics (e.g., BLEU, ROUGE) and clinical relevance, ensuring generated reports are both accurate and contextually meaningful. By systematically comparing LLM performance and prompt impacts, we provide a robust methodology for evaluating natural language generation (NLG) systems in radiology, setting a new standard for automated medical report generation. Our contributions to this paper are stated below:

1. We evaluate the semantic performance of four distinct large LLMs–Llama 3.2, Mistral 7B, Phi 3 Medium, and Gemma 2–by systematically comparing their ability to analyze the semantic relationships between generated reports from the R2GenRL model.
2. We integrate prompt engineering into the score evaluation framework, implementing three distinct prompting techniques to assess their influence on LLM performance and the score evaluation mechanism, thereby providing insights into the role of prompts in medical report generation.

3. We propose the use of semantic scoring (on a scale of 0–10) as a novel quality metric, leveraging LLMs and prompt engineering to bridge the gap between syntactic and contextual evaluation, ensuring that the generated reports are both semantically accurate and contextually relevant.
4. We conduct a comprehensive multi-model comparison within the specialized domain of radiology, emphasizing the performance of LLMs in the context of medical report generation, which serves as a foundation for informed model selection in healthcare applications.

2 Literature Review

Automated radiology report generation has emerged as a critical task in medical artificial intelligence, driven by the need to streamline radiological workflows and reduce the burden on clinicians. Four key research areas collectively address the challenges of generating and evaluating radiology reports. These areas include Generative AI and Multimodal Models, Natural Language Generation (NLG) in radiology, evaluation metrics and LLMs in Medical Natural Language Processing (NLP).

2.1 Generative AI and Multimodal Models

Generative AI, a rapidly evolving field, encompasses models that produce high-quality, human-like content, including text, images, and multimodal outputs. The AI research community often focuses on complex generative models that create realistic outputs, though the term "Generative AI" lacks a universal definition, leading to varied interpretations across domains [20]. In medical imaging, generative AI has driven advancements in radiology report generation and image analysis. The growth of generative AI for synthetic multimedia content, such as images and videos, emphasizing techniques like diffusion models and the need for robust datasets to support multimodal tasks [21]. Over the years, various models [22–25] have been proposed to generate structured, coherent, and clinically meaningful reports from radiographic images. While these models, such as R2GEN [26] and its variants [22,27–29] leverage encoder-decoder architectures to translate radiographic images into textual reports, integrating visual and textual information. Complementary approaches, such as contrastive learning with Momentum Contrast (MoCo) and ResNet backbones, enhance feature extraction for chest X-ray classification, particularly in data-scarce settings [30]. Bougueffa et al. (2024) underscore the importance of datasets for training such models and the potential of multimodal generative AI to advance medical imaging tasks [21]. Additionally, preprocessing techniques like Reinhard and Macenko normalization improve image quality, further boosting model performance [31]. This area appeals to AI and medical imaging researchers exploring multimodal generative systems.

2.2 Natural Language Generation (NLG) in Radiology

NLG focuses on producing coherent, contextually relevant text from structured or unstructured data, a critical capability for radiology report generation. Models like R2GenRL [22] combine visual feature extraction with language generation to produce structured reports from chest X-rays. However, generating reports that capture complex medical semantics remains challenging, as slight phrasing differences can alter clinical meaning. Traditional NLG evaluation metrics—BLEU [6], ROUGE [8], METEOR [7], and CIDEr [9]—prioritize lexical overlap, often failing to assess semantic accuracy or clinical relevance. This limitation has drawn attention from NLP and medical informatics researchers seeking evaluation methods that align with the diagnostic needs of radiology.

2.3 Semantic Evaluation Metrics

Evaluating radiology reports requires metrics that prioritize semantic alignment over surface-level similarity. Early approaches used static embeddings like Word2Vec [32] and GloVe [33] to measure conceptual

similarity. More recently, transformer-based models, such as BERT [34], BERTScore [35], BLEURT [36], and CLIPScore [37], have enabled more nuanced semantic comparisons. For instance, Phan et al. (2024) [38] employed BERTScore and CLIPScore to assess medical image captions in the ImageCLEFmedical 2024 challenge, highlighting their effectiveness for multimodal tasks. Raj et al. (2023) [39] proposed the Ask-to-Choose (A2C) prompting strategy to improve semantic consistency in text generation, addressing limitations of lexical metrics. These advancements attract researchers in NLP and clinical informatics focused on developing robust evaluation frameworks for medical texts.

2.4 Large Language Models in Medical NLP

LLMs, such as GPT-based models [40–42], have revolutionized NLP by capturing complex contextual relationships and generating human-like text. Thirunavukarasu et al. (2023) emphasized the potential of LLMs like ChatGPT and Med-PaLM 2 in clinical settings, noting their ability to process domain-specific text (e.g., patient notes, medical guidelines) and improve performance in biomedical NLP tasks [43]. In radiology, LLMs excel in zero-shot or few-shot settings, enabling flexible evaluation of generated reports without extensive fine-tuning [44]. For example, Doshi et al. (2024) [45] demonstrated that LLMs like ChatGPT-4 and Gemini can simplify radiology report impressions, with prompt design significantly influencing performance. However, challenges such as LLMs' struggles with conciseness and verisimilitude in medical summarization, underscoring the need for task-specific fine-tuning and human oversight in high-stakes settings [43]. Hu et al. (2024) [46] further highlighted limitations in LLMs' ability to summarize radiology reports accurately. These findings engage AI and medical NLP researchers exploring LLMs' role in clinical applications.

2.5 Research Gap and Technical Problem

Despite these advancements, significant gaps remain in evaluating the semantic accuracy of AI-generated radiology reports. Traditional metrics like BLEU and ROUGE fail to capture clinical relevance, while semantic evaluation methods often lack the flexibility and domain-specific understanding required for medical texts. By incorporating domain-specific text and task-specific fine-tuning LLMs performance can be enhanced, but current approaches rarely integrate these with systematic prompt engineering [43]. Moreover, while models like R2GenRL excel in report generation, their evaluation has relied on limited metrics or general-purpose models, hindering clinical adoption. Our work addresses this gap by proposing a novel LLM-based framework that leverages R2GenRL-generated reports and LLMs, and employs carefully designed prompt templates to assign semantic similarity scores (0–10) between generated and ground truth reports. Our framework proposed the role of LLMs in semantic evaluation, advancing the reliability of AI-driven medical text generation.

3 Methodology

This section outlines the proposed systematic approach used to evaluate the semantic meaning of the AI-generated report with the ground truth. We used the R2GenRL model to train the model on IU-Xray and predict the report on the test data of IU-Xray and MIMIC-CXR dataset, and by using large language models and prompt engineering techniques, we calculated the accuracy score based on semantic alignment. An overview of our proposed semantic score evaluation method is shown in Fig. 1.

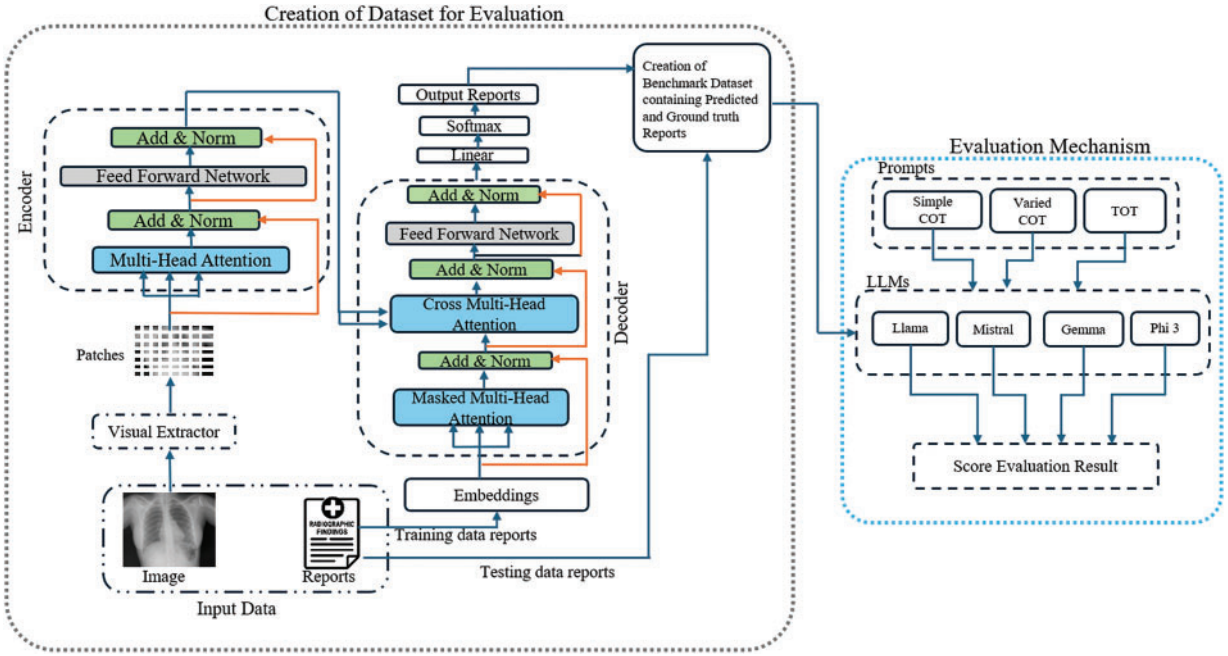

Figure 1: The overall architecture of our proposed semantic evaluation method, where the R2GenRL model in a simplified version is shown in gray dash boxes, while the Evaluation method is illustrated in blue dash line

3.1 Data Preprocessing

In the data preprocessing step, the document was first converted to lowercase. After converting the text, the reports were tokenized. Furthermore, irrelevant or extraneous characters, such as special symbols, were removed. After cleaning, the dataset was split into train (70%), validation (15%) and test (15%) sets.

3.2 Report Generation

Automatic report generation is quite a hectic task. We used a model (a reinforced cross-modal alignment approach) that was already built by [17]. This model is also known as R2GenRL. We trained this model on the IU-Xray dataset by changing different parameters, specifically the report's length. By changing the parameters like token length and evaluating the result using natural language generation (NLG) metrics, we obtained a Bleu-4 score of 26.3 for the test data, which is quite good compared to other models, as shown in Table 1.

Table 1: Performance comparison of different existing methods on the test set of IU-Xray dataset with respect to NLG metrics

Method	Bleu1	Bleu2	Bleu3	Bleu4	Rouge_L
R2GenCMN [22]	0.470	0.304	0.219	0.165	0.371
SentSAT+KG [47]	0.441	0.291	0.203	0.147	0.367
XPRONET [24]	**0.525**	0.357	0.262	0.199	0.411
KiUT [48]	**0.525**	0.360	0.251	0.185	0.409
R2GenRL [17]	0.482	**0.365**	**0.302**	**0.263**	**0.451**

In the next step, we generated the report on the test data of the IU-Xray dataset and a small subset of MIMIC-CXR data using the above-trained model. After that, we created a benchmark dataset by combining the generated AI reports from the model with the ground truth of the IU-Xray test and MIMIC-CXR data. The benchmark dataset is used further in our semantic score evaluation mechanism.

3.3 Large Language Model Selection

Four cutting-edge LLMs, i.e., Llama 3.2, Mistral, Phi 3 Medium, and Gemma 2, were selected for this study to introduce the semantic evaluation mechanism. In addition to this, the comparative performance of these four LLMs was also analyzed in this semantic score evaluation process. The primary reason for choosing these models for the evaluation process is because of their open-source availability. Unlike proprietary LLMs, such as OpenAI's GPT series, the selected models allow for full transparency and adaptability. Open-source models facilitate fine-tuning, making them more practical for specific applications like medical report generation. Furthermore, their open availability ensures the reproducibility of results, a cornerstone of scientific research. While GPT-4 and similar proprietary models demonstrate advanced linguistic capabilities, their restricted access, and reliance on paid APIs make them less suitable for academic investigations, particularly in healthcare contexts where reproducibility is paramount. The Table 2 shows the parameters of the LLM model we used in our method.

Table 2: LLM models and their parameters used in our methodology

Model	Parameters (Billion)	Size of model (GB)
Llama 3.2	3 B	2.0 GB
Mistral	7 B	4.1 GB
Phi 3 medium	14 B	7.9 GB
Gemma 2	9 B	5.5 GB

3.4 Prompt Engineering

Prompt design plays a pivotal role in determining the quality and accuracy of the generated text. The key features of these prompts we used in our methodology are:

1. **Role Definition:** AI is instructed to adopt the role of an expert medical evaluator.
2. **Input Context:** The comparison is between two specific texts: the Ground Truth (reference) and the Predicted Report.
3. **Evaluation Focus:** Specifies clinical areas of focus, such as "lung volumes," "pleural effusion," "heart size," and "lung markings."
4. **Scoring Criteria:** Provides a clear rubric for assigning a score (0–10) based on the level of similarity and detail accuracy.
5. **Output Format:** Enforces a structured response format: [Score: Obtained Score/10, Reason: explanation].
6. **Context-Specific Guidance:** Ensures relevance to the medical domain, emphasizing key clinical findings over peripheral content.

Three distinct prompts were developed to assess how variations in prompt design influence the performance of the models:

1. **Zero-Shot Prompt:** This prompt adopts a direct approach where the model is instructed to compare Ground Truth and Predicted Reports based on key clinical findings such as "lung volumes," "pleural effusion," and "heart size." It guides the model to evaluate the semantic similarity and assign a score (0–10) based on a clear rubric, ensuring a concise and structured output format.
2. **Chain of Thought Prompt:** This version introduces a more detailed internal reasoning framework, breaking the evaluation process into stages such as key element analysis, critical differences assessment, and impact evaluation. While maintaining the output simplicity, it enhances the robustness and consistency of the model's reasoning process across different scenarios.
3. **Tree of Thoughts Prompt:** This advanced prompt structures the evaluation into hierarchical branches, focusing on structural analysis, clinical findings, and implications. Each branch assesses specific aspects like terminology usage, clinical alignment, and treatment impact, with scores synthesized into a weighted final evaluation. This method captures nuanced details while prioritizing clinically significant findings.

Defining Semantic Scoring

In this study, the semantic scoring is defined as the degree to which the AI-generated report accurately captures the clinical findings of the ground truth report, as assessed by the LLMs on a 0–10 scale. The precision emphasizes semantic alignment and prioritizes the inclusion of critical diagnostic features like lung volumes, heart size, pleural effusion etc., over lexical overlap. For example:

- A perfect score of **10** was given when the generated report described *clear lungs, no infiltrates, and a normal cardiac silhouette*, matching the ground truth exactly in clinical content.
- A low score of **2** was assigned when the ground truth noted *pulmonary edema and interstitial disease*, but the model generated findings like *low lung volumes and bibasilar disease*.

These examples illustrate how omission or misrepresentation of critical findings impacts semantic alignment. Additionally, several quantitative metrics were used to support these semantic evaluations. BERTScore-Precision (0.877), which measures token-level embedding similarity, serves as a complementary metric to LLM-based semantic precision.

Each model was tested using all three prompts, and the outputs were analyzed to identify prompt sensitivity. In the same way, the semantic score of each predicted report is compared with the ground truth report, and the score is calculated using the LLM and prompts.

4 Experiment

The proposed semantic score evaluation mechanism for medical report generation was tested on the IU–Xray dataset. Furthermore, we used NVIDIA RTX A5000 24 GB to train the report generation model. For the semantic score evaluation process, where we implemented different LLM models, we used the free resource of Google Colab GPU.

Dataset

The dataset we used for test purpose in this frame work are IU-Xray and MIMIC-CXR. These datasets were selected due to their widespread adoptionin the field of medical report generation. According to the survey presented in [49], the IU X-Ray dataset has been utilized in over 71 research papers, while MIMIC-CXR has been used in 68 studies. This high frequency of use reflects their credibility, diversity, and relevance, making them suitable benchmarks for evaluating the effectiveness of medical report generation models. IU X-ray includes 3995 reports with 7470 chest X-ray images, split into 70% training (~5229 images), 15% validation (~1120 images), and 15% testing (~1120 images), generating 1117 benchmark reports. As we have

trained R2GENRL model on the IU-Xray dataset, so the large portion of the dataset was used in training and evaluation, and only 1117 reports were limited to use for the the semantic evaluation framework. In the same way, MIMIC-CXR contains around 227,835 reports with around 377,110 images. In our case to standardize the practice and evaluate the result, we randomly used 1120 test reports (~1120 images) sampled for evaluation from the MIMIC-CXR, so that both dataset must contain the same number of reports. The selection of a limited number of data also helped us in evaluating the semantic scores using the different models, as we used the free version of the Google Colab GPU. As there is some limitation of the Colab Free tier, like idle timeout, as if the system remain inactive (30–90 min), the session will automatically disconnect. So due to computational complexity, the choice of limited data was also a good choice.

5 Results and Analysis

This analysis was formulated to assess the LLM model's behavior, evaluate the semantic score on a scale of 0 to 10, and understand how different LLM models perform based on different prompting techniques. Figs. 2–4 show the experimental results of different LLM models with different prompts on the created benchmark dataset. The score distribution across the figures was derived from the semantic similarity calculations from the predicted and ground truth reports. These figures reflect the performance of the Llama 3.2, Mistral, Phi 3 Medium, and Gemma 2 models. Mistral model demonstration in the whole process indicates a high degree of semantic alignment and consistency between the reports, with scores predominantly concentrated at the maximum value of 10. Similarly, the Phi 3 Medium model, with most scores clustering between 9.0 and 10.0, reflects robust semantic matching capabilities. In contrast, the Gemma 2 model displays a slightly broader range of scores, typically centered between 8.0 and 9.0, with somewhat less optimal semantic alignment than Mistral and Phi 3 Medium. The Llama 3.2 model, however, exhibits a peak around 8.0 with more variability, suggesting a less consistent semantic similarity between the predicted and ground truth reports. Some of the semantic scores significantly deviating from the typical ranges (e.g., Gemma 2: 8.0–9.0, Llama 3.2: 8.0 with broader spread, as shown in Figs. 2–4), were included in the analysis without removal to preserve the integrity of the 1117-report benchmark dataset. The broader score distributions for Gemma 2 and Llama 3.2 (e.g., lower tails in Fig. 2a for Llama 3.2) indicate less consistent semantic alignment, potentially due to sensitivity to prompt variations or lack of radiology-specific fine-tuning. Mean scores such as 0.768 for Gemma 2 and 0.801 for Llama 3.2 under the TOT prompting approach incorporate these outliers, reflecting overall model performance. No statistical outlier detection methods (e.g., interquartile range) were applied, as the study prioritized raw score distributions to capture model variability.

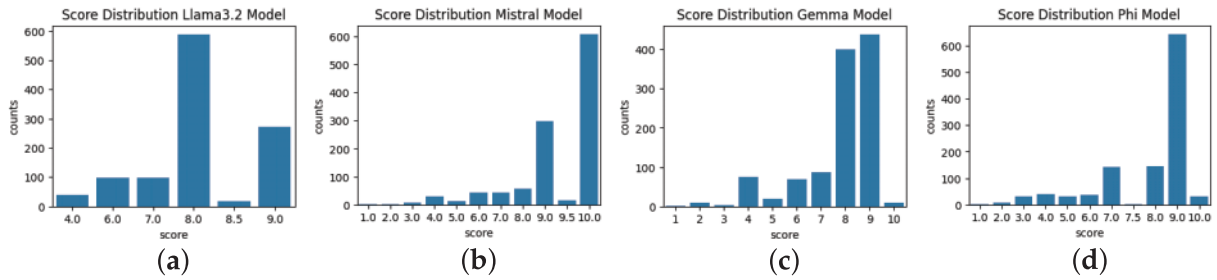

Figure 2: Semantic score distribution (0–10) of AI-generated reports using zero-shot prompting: (**a**) Llama 3.2 shows high variance, (**b**) Mistral and (**c**) Phi 3 peak at 10, (**d**) Gemma 2 shows moderate consistency

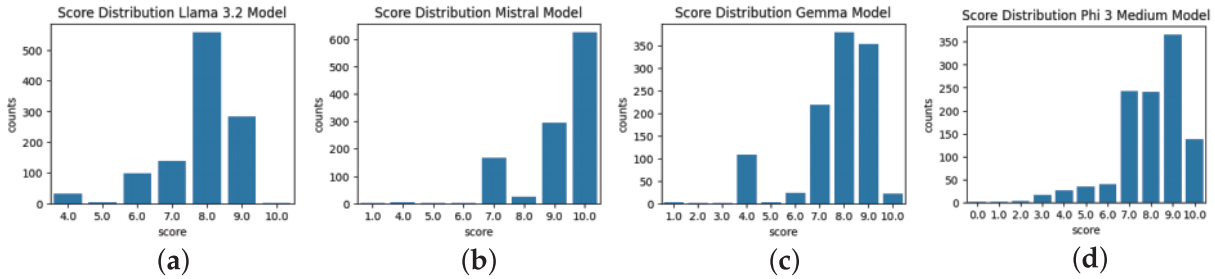

Figure 3: Semantic score distribution using Chain of Thought prompting. (**b**) Mistral, (**c**) Gemma 2, and (**d**) Phi 3 Medium show significant clustering near maximum scores (9.0–10.0), demonstrating superior semantic alignment compared to (**a**) Llama 3.2

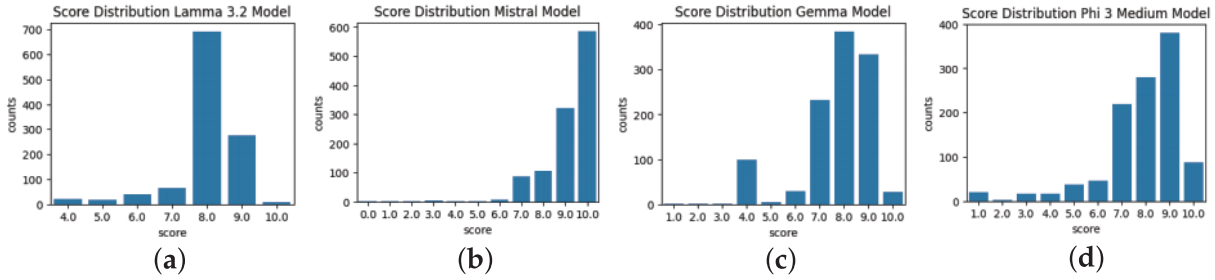

Figure 4: Score distributions under Tree of Thoughts prompting. All models shift toward higher scores compared to Fig. 3, with (**a**) Llama 3.2 peaking at 8–9, (**b**) Mistral and (**d**) Phi 3 Medium showing marked improvements, and (**c**) Gemma displaying a more balanced shift in semantic alignment

In addition, Fig. 5 presents a comparative analysis of four different language models, Llama 3.2, Mistral, Gemma 2, and Phi 3 Medium, with respect to three prompting strategies: Zero-Shot, Chain of Thought (COT), and Tree of Thoughts (TOT). The results highlight that Mistral has consistently outshone all other models, scoring the maximum in all prompting strategies with 86.71% in Zero-Shot, 90.02% in COT, and 91.97% in TOT. Llama 3.2 consistently improved with the prompts from 74.41% in Zero-Shot to 78.30% in COT and finally 80.13% in TOT. While Gemma 2 and Phi 3 Medium showed similar performance trends, they lagged Mistral and Llama 3.2. The improvement in performance levels for COT and TOT means that some form of structured reasoning is aiding the responses of a model, and the improvements seem to be the most substantial with Mistral. Note that while it appears all models become more capable of detailed prompts, the differences in performance between COT and TOT have not been very pronounced, meaning that the extra reasoning structure from switching from sequential to tree-based does not provide substantial extra value.

Similarly, Fig. 6 reorganizes the results from the perspective of putting them in groups based on models rather than by prompting strategies, thus illustrating the relative strength of each model across all prompts. From these, Mistral appears to have improved more by using structured reasoning, with a great jump from Zero-Shot to COT and a slight increase from COT to TOT. Llama 3.2 has also improved significantly with advanced prompting, while Gemma 2 and Phi 3 Medium appear to have improved rather slightly. Interestingly, it is worth mentioning that Phi 3 Medium declined mildly from COT (81.91%) to TOT (80.62%), which suggests that tree-based reasoning may not be all that helpful occasionally, depending on

the architectural set-up of the model. Across all four models, COT and TOT are inarguably better than Zero-Shot, thus further stacking credence on the strength of structured prompting strategies for large language model performance.

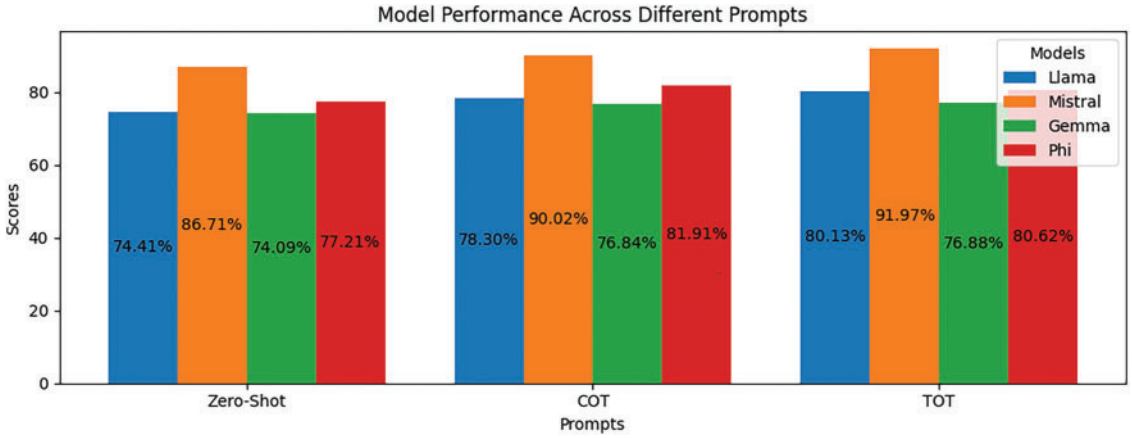

Figure 5: Performance comparison of models (Llama 3.2, Mistral, Gemma 2, Phi 3 Medium) across three prompting strategies: zero-shot, Chain of Thoughts (COT), and Tree of Thoughts (TOT)

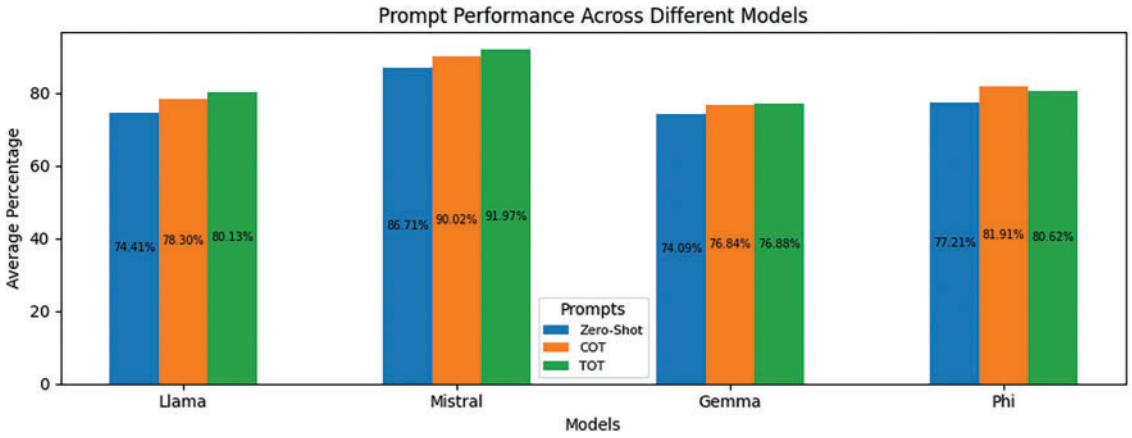

Figure 6: The average performance of three prompting strategies, Zero-Shot, Chain of Thoughts (COT), and Tree of Thoughts (TOT), across different LLMs (Llama 3.2, Mistral, Gemma 2, Phi 3 Medium)

To further validate our findings, we evaluated the models using the MIMIC-CXR dataset. The Table 3 presents the semantic scores for different prompting strategies. Although the scores on MIMIC-CXR are slightly lower. For instance, Mistral achieved approximately 90.0% in the Tree of Thoughts (ToT) prompt, this is likely due to the increased variability and complexity of findings in the MIMIC-CXR dataset. Nevertheless, the overall performance ranking remained consistent, with Mistral performing the best, followed by Phi 3 Medium, Gemma 2, and Llama 3.2.

To evaluate and understand the rationale behind the scores assigned by LLM models to predicted reports compared to the ground truth, we also generate explanations alongside the respective scores. This approach provides insights into the model's decision-making process and helps assess the alignment between predicted

outputs and reference reports. Table 4 shows a few examples of the reasons behind the scores assigned to reports.

Table 3: Semantic similarity scores on UI-Xray and MIMIC-CXR dataset across prompting strategies

Model	DataSet	Zero shot	COT	TOT
Llama 3.2	IU X-ray	74.4	78.3	80.1
Llama 3.2	MIMIC-CXR	73.0	77.0	79.0
Mistral 7B	IU X-ray	86.7	90.0	91.9
Mistral 7B	MIMIC-CXR	85.0	88.0	90.0
Gemma 2	IU X-ray	74.0	76.8	76.8
Gemma 2	MIMIC-CXR	73.0	75.0	75.0
Phi 3 medium	IU X-ray	77.2	81.9	80.6
Phi 3 medium	MIMIC-CXR	76.0	80.0	79.0

Table 4: Performance comparison of LLMs with reasons for assigned scores on radiological report alignment (Few examples)

Ground truth (Summary)	Predicted (Summary)	Score (Out of 10)	Reason by LLM (Summary)	LLM model
Normal chest: clear lungs, no abnormality.	Also clear lungs, normal heart and bones.	9	Key findings match; minor phrasing differences.	Llama 3.2
Findings include calcified granuloma in lung.	No mention of granuloma.	4	Misses important lung detail (granuloma).	Llama 3.2
Clear lungs, no infiltrates, normal silhouette.	Similar findings, minor detail differences.	10	Fully aligned in meaning despite minor omissions.	Mistral
Emphysema, patchy opacities in lower lobe.	Cardiomegaly and edema; different focus.	3	Important findings differ; focus mismatch.	Mistral
Cardiomegaly with pulmonary edema over chronic interstitial disease.	Low lung volumes with bibasilar disease and bronchovascular crowding.	2	Major findings differ: edema and interstitial changes in GT vs. low volumes in prediction.	Phi 3 Medium
No acute findings; normal heart and lungs. Notes surgical clips.	No acute findings; heart and lungs normal. Surgical clips not mentioned.	9	All major findings match; minor omission of surgical clips.	Gemma 2

Furthermore, Table 5 presents a comparative analysis of the performance of various models when evaluating the semantic score using different prompt techniques (zero-shot, COT, and TOT) on the same predicted report. This evaluation highlights the impact of prompt engineering, although the difference is not that much, as we didn't change the overall context in the prompt templates.

Tables 6 and 7 support our study that LLMs can be used as semantic evaluation metrics. Table 6 shows the semantic similarity scores (percentage accuracy) calculated by the LLMs when generated reports and ground truth reports were given to the LLMs. In the same way, Table 7 provides the accuracy of the established metrics on the same reports generated by the R2GenRL model. The LLMs' scores in Table 6, especially Mistral's (up to 91.9%) and Phi 3 medium's (up to 81.9%), are very close to or even exceed the BERTScore-F1 (88.1%) and CLIP-Score (88.7%), showing that these LLMs can evaluate semantic similarity as well as or better than traditional metrics. Even Llama 3.2 and Gemma 2, with lower scores (74.0% to

80.1%), are not far from the conventional metrics, indicating they are still effective. This close alignment with established metrics supports the proposal that LLMs can be reliable semantic evaluation metrics. However, the lower BLEURT score (39.6%) suggests they may need improvement in capturing fine-grained details.

Table 5: Performance comparison of LLMs using different prompt techniques on radiological report alignment (Few examples)

Ground truth	Predicted	Score				Prompt
		Llama 3.2	Mistral	Gemma 2	Phi 3 Medium	
Low lung volumes with increased lung markings particularly in the left perihilar region xxxx related to history of bronchitis. no acute infiltrate. the heart is normal in size. the mediastinum is within normal limits the lungs are hypoaerated. there is mild increase in perihilar markings xxxx related to patient's history bronchitis. no acute infiltrate or pleural effusion are seen.	No acute cardiopulmonary findings the cardiomediastinal silhouette and pulmonary vasulature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings.	8	9	7	5	Zero-Shot
		8	7	7	7	COT
		8	9	7	7	TOT
No radiographic evidence of acute cardiopulmonary disease heart xxxx mediastinum xxxx bony structures and lung xxxx are unremarkable. stable small calcified granuloma left base. no xxxx acute findingsopacitiesinfiltrates noted.	No acute cardiopulmonary abnormality. the lungs are clear bilaterally. specifically no evidence of focal consolidation pneumothorax or pleural effusion. cardio mediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality.	4	8	7	7	Zero-Shot
		6	9	6	8	COT
		7	9	8	7	TOT

Table 6: Performance metrics of various LLMs with different prompting techniques

Model	Llama 3.2			Mistral			Gemma 2			Phi 3 medium		
	Zero-Shot	COT	TOT	Zero-Shot	COT	TOT	Zero-Shot	COT	TOT	Zero-Shot	COT	TOT
Score	0.744	0.783	0.801	0.867	0.900	0.919	0.740	0.768	0.768	0.772	0.819	0.806

Table 7: Evaluation metrics for radiology report generation

BERTScore-Precision	BERTScore-Recall	BERTScore-F1	CLIP-Score	BLEURT
0.877	0.885	0.881	0.887	0.396

To further validate the results, we performed a statistical test to verify that the differences in semantic scores across models and prompt types are significant, not due to random variation. The paired t-test analysis revealed in Table 8 statistically significant differences in semantic scores between most model pairs. The statistical comparison of model performances revealed that Mistral significantly outperforms all other models, including Llama 3.2, Gemma 2, and Phi 3 Medium, as indicated by highly significant p-values ($p < 0.01$) in all related pairwise t-tests. This performance gap is likely due to Mistral's better adaptation to medical image-text tasks, as reflected in its higher semantic scores across both IU X-ray and MIMIC-CXR datasets. In contrast, no statistically significant differences were found between Llama 3.2, Gemma 2, and Phi 3 Medium ($p > 0.1$), suggesting comparable capabilities among these models. These findings support the conclusion that while lightweight models like Phi 3 Medium and Gemma 2 offer competitive baseline performance, Mistral provides a clear advantage for tasks demanding higher semantic accuracy in medical report generation.

Table 8: Statistical comparison of model performance

Model pair	t-statistic	p-value
Llama 3.2 vs. Mistral	−64.30	0.00024
Llama 3.2 vs. Gemma 2	2.05	0.17678
Llama 3.2 vs. Phi 3 Medium	0.885	0.13171
Mistral vs. Gemma 2	18.69	18.69
Mistral vs. Phi 3 Medium	10.40	0.00912
Gemma 2 vs. Phi 3 Medium	−7.19	0.01879

To assess the effect of prompt variation on model performance, we conducted an ANOVA test for each model, considering the three prompts we used in our framework. The results, as shown in the Table 9, reveal a statistically significant difference in performance across different prompts. For example, in the case of Llama 3.2 ($F = 26.80$, $p = 0.00102$) and Phi 3 Medium ($F = 23.92$, $p = 0.00138$) showed highly significant sensitivity to prompt variations, suggesting the effect of prompt-dependent behavior. Mistral also demonstrated a significant effect ($F = 10.64$, $p = 0.01065$), while Gemma 2 displayed a marginally significant difference ($F = 5.35$, $p = 0.04637$), indicating a weaker but still noticeable impact of prompt choice. These findings highlight the importance of prompt engineering when using LLMs for medical report result evaluation, as the model outputs are meaningfully influenced by the prompts.

Table 9: Statistical comparison of model performance using ANOVA across different prompts

Model	F-statistic	p-value
Llama 3.2	26.800	0.00102
Mistral	10.64	0.01065
Gemma 2	5.35	0.04637
Phi 3 Medium	23.92	0.00138

While overall performance was promising, a detailed failure case analysis was performed to uncover critical weaknesses in the LLMs' scoring behavior. We examined cases where LLMs assigned misleading scores, revealing critical limitations as summarized in Table 10. Several examples illustrate common error

patterns. For instance, one report received a high score of 9/10 despite omitting the clinically significant mention of "mildly diminished lung volumes." In another case, a report scored 8/10 even though it failed to capture key chronic findings such as emphysema and granuloma. Additionally, one model assigned a high score based largely on lexical overlap with the reference report, despite introducing a false positive observation, such as an inaccurately enlarged heart. These errors can be seen primarily because of excessive dependence on surface-level similarity, and prompt ambiguity. These findings highlight the need for improved prompt engineering or model fine-tuning to better align automated scoring with clinical accuracy.

Table 10: Failure cases highlighting common error patterns in LLM-generated clinical report scores

Ground truth	Predicted	Predicted score
No acute cardiopulmonary findings. Mildly diminished lung volumes. Normal cardiac silhouette, mediastinal contour, pulmonary vasculature. No focal consolidation, pleural effusion.	No acute cardiopulmonary abnormality. Lungs clear bilaterally. Normal cardiac silhouette. No focal consolidation, pneumothorax, or pleural effusion. Mediastinal contours unremarkable.	9/10 (Mistral)
Hyperexpanded lungs consistent with emphysema. No active disease. Normal heart size, pulmonary vascular. Calcified lymph and granuloma noted. No focal airspace disease, pleural effusion, or pneumothorax.	No acute cardiopulmonary findings. Lungs clear bilaterally. Normal heart size. No focal consolidation, pneumothorax, or pleural effusion. Mediastinal contours unremarkable.	8/10 (Phi 3 Medium)
No acute cardiopulmonary process. Normal cardiomediastinal silhouette. Lungs clear without consolidation. No pneumothorax, large pleural effusion, or bone abnormality.	No acute cardiopulmonary findings. Lungs clear bilaterally. Heart size slightly enlarged. No focal consolidation, pneumothorax, or pleural effusion. Normal mediastinal contours.	8/10 (Llama 3.2)
Pulmonary vascular congestion, emphysema, bibasilar streaky opacities, borderline enlarged heart, aortic calcifications, hyperinflated lungs, chronic interstitial markings. No focal airspace disease or acute bony abnormality.	No acute cardiopulmonary findings. Hyperinflated lungs consistent with emphysema. Normal heart size. No focal consolidation, pneumothorax, or pleural effusion. Normal mediastinal contours.	7/10 (Phi 3 Medium)

6 Discussion

This study presents a novel framework for evaluating the semantic accuracy of AI-generated radiology reports, addressing the shortcomings of traditional lexical metrics like BLEU and ROUGE, which often overlook clinical relevance. By employing large language models (LLMs) and advanced prompt engineering–Zero-Shot, Chain of Thought (COT), and Tree of Thoughts (TOT)—we developed a semantic scoring system (0–10 scale) prioritizing diagnostic accuracy. Experiments on IU X-ray and MIMIC-CXR datasets show Mistral 7B, especially with TOT prompting, achieving superior semantic alignment (91.9%), outperforming Llama 3.2, Phi 3 Medium, and Gemma 2. Statistical analyses, including paired t-tests ($p < 0.01$) and

ANOVA ($p < 0.05$), confirm the significant impact of structured prompting. Our semantic scoring correlates closely with BERTScore-F1 (88.1%) and CLIP-Score (88.7%), surpassing lexical metrics by evaluating clinical equivalence. However, failure cases, reveal issues like over-reliance on lexical overlap, as seen in inflated scores for reports missing key findings (e.g., "emphysema"), highlighting the need for radiology-specific fine-tuning. Limitations include reliance on small portion of test data, potentially limiting generalizability, and lack of radiologist validation. Additionally, computational constraints from Google Colab's free-tier GPU also restricted experiment scale. Future work should diversify datasets, integrate expert feedback, and explore dynamic prompting. Fine-tuning on radiology corpora and incorporating multimodal inputs, could enhance performance. Lightweight LLMs may improve scalability for clinical use. This framework advances medical NLP by offering a clinically relevant evaluation metric, setting a foundation for reliable AI-driven radiology systems.

7 Conclusion

In this paper, we introduced a novel approach for evaluating the semantic accuracy of the AI-generated radiology reports using LLMs and advanced prompt engineering techniques. Using a semantic scoring value (0–10), this framework focused on diagnostic relevance, which addresses the limitations of traditional evaluation methods by prioritizing clinical correctness over surface-level text similarity. Experimental results on the IU X-ray and MIMIC-CXR datasets show that Mistral 7B, particularly when guided by the Tree of Thoughts (ToT) prompting strategy, achieves superior semantic alignment scores (up to 91.9%), outperforming Llama 3.2, Phi 3 Medium, and Gemma 2. These findings are further supported by strong correlation with established metrics such as BERTScore-F1 (88.1%) and CLIP-Score (88.7%). The statistical analyses, including paired t-tests ($p < 0.01$) and ANOVA (e.g., $p = 0.00102$ for Llama 3.2), also demonstrate the significant impact of prompt engineering, with TOT and COT strategies enhancing performance across all models. Despite these improvements, failure cases reveal persistent challenges such as over-reliance on lexical overlap, and prompt ambiguity, highlighting the ongoing need for domain-specific fine-tuning. These results underscore the potential of LLMs, when appropriately guided, enhances the reliability and clinical validity of automated radiology report generation, and set a foundation for more trustworthy and context-aware AI systems in medical imaging.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00436773).

Author Contributions: Haider Ali: methodology, conceptualization, writing—original draft preparation, formatting. Rashadul Islam Sumon: writing—original draft preparation, data curation. Abdul Rehman Khalid: data curation, writing—original draft preparation, visualization. Kounen Fathima: data curation, visualization. Hee Cheol Kim: supervision, project administration, funding acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The widely recognized ROCO dataset, which is publicly accessible, was utilized.

Ethics Approval: Not applicable

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Hou WJ, Cheng Y, Xu KS, Li WJ, Liu J. RECAP: towards precise radiology report generation via dynamic disease progression reasoning. arXiv:2310.13864. 2023.
2. Plesner LL, Mü ller FC, Nybing JD, Laustrup LC, Rasmussen F, Nielsen OW et al. Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology. 2023;307(3):e222268. doi:10.1148/radiol.222268.
3. Sloan P, Clatworthy P, Simpson E, Mirmehdi M. Automated radiology report generation: a review of recent advances. IEEE Rev Biomed Eng. 2024;307(3):e222268. doi:10.1109/RBME.2024.3408456.
4. Zhang L, Wen X, Li JW, Jiang X, Yang XF, Li M. Diagnostic error and bias in the department of radiology: a pictorial essay. Insights Imaging. 2023;14(1):163. doi:10.1186/s13244-023-01521-7.
5. Palmer J. Report finds radiologists to blame for missed diagnoses. Patient Safety & Quality Healthcare (PSQH); 2028 Dec 4. [Online]. [Accessed on 2025 Jun 10]. Available from: https://www.psqh.com/analysis/report-finds-radiologists-to-blame-for-missed-diagnoses-2/.
6. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA; 2002. p. 311–8. doi:10.3115/1073083.1073135.
7. 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. Ann Arbor, Michigan; 2005. p. 65–72.
8. Lin CY. Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81.
9. 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. Boston, MA, USA; 2015. p. 4566–75.
10. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv:2302.13971. 2023. doi:10.48550/arXiv.2302.13971.
11. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv:2310.06825. 2023. doi:10.48550/arXiv.2310.06825.
12. Team G, Riviere M, Pathak S, Sessa PG, Hardin C, Bhupatiraju S, et al. Gemma 2: improving open language models at a practical size. arXiv:2408.00118. 2024. doi:10.48550/arXiv.2408.00118.
13. Abdin M, Aneja J, Awadalla H, Awadallah A, Awan AA, Bach N, et al. Phi-3 technical report: a highly capable language model locally on your phone. arXiv:2404.14219. 2024. doi:10.48550/arXiv.2404.14219.
14. Li MC, Huang JT, Yeung J, Blaes A, Johnson S, Liu HF, et al. CancerLLM: a large language model in cancer domain. arXiv:2406.10459. 2024. doi:10.48550/arXiv.2406.10459.
15. Habchi Y, Kheddar H, Himeur Y, Belouchrani A, Serpedin E, Khelifi F, et al. Advanced deep learning and large language models: comprehensive insights for cancer detection. Image Vis Comput. 2025;157:105495. doi:10.1016/j.imavis.2025.105495.
16. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: techniques and applications. arXiv:2402.07927. 2024. doi: 10.48550/arXiv.2402.07927.
17. Qin H, Song Y. Reinforced cross-modal alignment for radiology report generation. In: Findings of the association for computational linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 448–58. doi:10.18653/v1/2022.findings-acl.38.
18. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inform Process Syst. 2022;35:24824–37. doi:10.48550/arXiv.2201.11903.
19. Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. Adv Neural Inf Process Syst. 2024;36. doi:10.48550/arXiv.2305.10601.
20. García-Peñalvo F, Vázquez-Ingelmo A. What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. Int J Interact Multimed Artif Intell (IJIMAI). 2023;8(4):7–16. doi:10.9781/ijimai.2023.07.006.

21. Bougueffa H, Keita M, Hamidouche W, Taleb-Ahmed A, Liz-López H, Martín A, et al. Advances in AI-generated images and videos. Int J Interact Multimed Artif Intell. 2024;9(1):1–36. doi:10.9781/ijimai.2024.11.003.
22. Chen Z, Shen Y, Song Y, Wan X. Cross-modal memory networks for radiology report generation. arXiv:2204.13258. 2022. doi: 10.48550/arXiv.2204.13258.
23. 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; 2021 Jun. p. 13753–62. doi:10.48550/arXiv.2106.06963.
24. Wang J, Bhalerao A, He Y. Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland; 2022. p. 563–79. doi:10.48550/arXiv.2207.04818.
25. Liu F, Yin C, Wu X, Ge S, Zou Y, Zhang P, et al. Contrastive attention for automatic chest X-ray report generation. arXiv:2106.06965. 2021. doi:10.48550/arXiv.2106.06965.
26. Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. arXiv:2010.16056. 2020. doi: 10.48550/arXiv.2010.16056.
27. Wang Z, Liu L, Wang L, Zhou L. R2genGPT: radiology report generation with frozen LLMs. Meta-Radiol. 2023;1(3):100033. doi:10.1016/j.metrad.2023.100033.
28. Wang X, Li Y, Wang F, Wang S, Li C, Jiang B. R2genCSR: retrieving context samples for large language model based X-ray medical report generation. arXiv:2408.09743. 2024. doi:10.48550/arXiv.2408.09743.
29. Sun Y, Lee YZ, Woodard GA, Zhu H, Lian C, Liu M. R2Gen-Mamba: a selective state space model for radiology report generation. arXiv:2410.18135. 2024. doi:10.48550/arXiv.2410.18135.
30. Vo TT, Do TN. Improving chest X-ray image classification via integration of self-supervised learning and machine learning algorithms. Korea Instit Inform Commun Eng. 2024;22(2):165–71. doi:10.56977/jicce.2024.22.2.165.
31. Rabeya RA, Bhattacharjee S, Kim D, Kim HC, Cho NH, Choi HK. An experimental comparison and quantitative analysis on conventional stain normalization for histopathology images. J Korea Multimedia Soc. 2024;27(11):1268–88.
32. Ma L, Zhang Y. Using Word2Vec to process big text data. In: 2015 IEEE International Conference on Big Data (Big Data). Santa Clara, CA, USA: IEEE; 2015. p. 2895–7. doi:10.1109/BigData.2015.7364114.
33. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar; 2014. p. 1532–43. doi:10.3115/v1/D14-1162.
34. Kenton JDMWC, Toutanova LK. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Minneapolis, Minnesota; 2019. Vol. 1, p. 4171–86. doi:10.18653/v1/N19-1423.
35. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: evaluating text generation with bert. arXiv:1904.09675. 2019. doi:10.48550/arXiv.1904.09675.
36. Sellam T, Das D, Parikh AP. BLEURT: learning robust metrics for text generation. arXiv:2004.04696. 2020. doi: 10.48550/arXiv.2004.04696.
37. Hessel J, Holtzman A, Forbes M, Bras RL, Choi Y. Clipscore: a reference-free evaluation metric for image captioning. arXiv:2104.08718. 2021. doi:10.48550/arXiv.2104.08718.
38. Phan TV, Nguyen TK, Hoang QA, Phan QT, Nguyen-Tat TB. UIT-2Q2T at ImageCLEFmedical 2024 caption: multimodal medical image captioning using bootstrapping language-image pre-training. In: CLEF 2024: Conference and Labs of the Evaluation Forum; 2024 Sep 9–12;. Grenoble, France; 2024.
39. Raj H, Gupta V, Rosati D, Majumdar S. Semantic consistency for assuring reliability of large language models. arXiv:2308.09138. 2023. doi:10.48550/arXiv.2308.09138.
40. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach. 2020;30:681–94. doi:10.1007/s11023-020-09548-1.
41. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inform Process Syst. 2022;35:27730–44. doi:10.48550/arXiv.2203.02155.

42. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv:2303.08774. 2023. doi:10.48550/arXiv.2303.08774.
43. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. doi:10.1038/s41591-023-02448-8.
44. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inform Process Syst. 2020;33:1877–901. doi:10.48550/arXiv.2005.14165.
45. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. Radiology. 2024;310(3):e231593. doi:10.1148/radiol.231593.
46. Hu D, Zhang S, Liu Q, Zhu X, Liu B. The current status of large language models in summarizing radiology report impressions. arXiv:2406.02134. 2024. doi:10.48550/arXiv.2406.02134.
47. Zhang Y, Wang X, Xu Z, Yu Q, Yuille A, Xu D. When radiology report generation meets knowledge graph. Proc AAAI Conf Artif Intell. 2020;34:12910–7. doi:10.1609/aaai.v34i07.6989.
48. Huang Z, Zhang X, Zhang S. KiUT: knowledge-injected U-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun. p. 19809–18. doi:10.48550/arXiv.2306.11345.
49. Xinyi W, Figueredo G, Li R, Zhang WE, Chen W, Chen X. A survey of deep-learning-based radiology report generation using multimodal inputs. Med Image Anal. 2025:103627.