RadEval: A novel semantic evaluation framework for radiology report

Hilya Tsaniya [a,1], Chastine Fatichah [a,2,*], Nanik Suciati [a,3]

[a] Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1] 7025222004@student.its.ac.id; [2] chastine@if.its.ac.id; [3] nanik@if.its.ac.id
* corresponding author

ARTICLE INFO

Article history
Received July 8, 2025
Revised October 10, 2025
Accepted October 11, 2025
Available online November 30, 2025

Keywords
Radiology report generation
Semantic evaluation
Clinical NLP
Medical ontology
Clustering

ABSTRACT

The evaluation of automatically generated radiology reports remains a critical challenge, as conventional metrics fail to capture the semantic, clinical, and contextual correctness required for automatic medical analysis. This study proposes RadEval, a semantic-aware evaluation framework, to assess the quality of generated radiology reports. This method integrates domain-specific knowledge and contextual embeddings to evaluate the quality of generated radiology reports using a four-level scoring system. Given a reference report and a predicted report from a radiology image, RadEval performs scoring evaluation by first extracting relevant medical entities using a fine-tuned biomedical NER model. These entities are normalized through ontology mapping using RadLex concept identifiers to resolve lexical variation. Then, semantically related entities were clustered using BioBERT's contextual embeddings to capture deeper semantic similarity. In addition, predicted abnormality tags are incorporated to weight clinically significant terms during score aggregation. The final semantic score reflects a weighted combination of exact match, ontology match, and contextual similarity, modulated by tag importance. Experiments were conducted on the MIMIC-CXR dataset, which contains over 200,000 report pairs. Comparative evaluations show that RadEval outperforms traditional metrics, achieving an F1-score of 0.69, compared to 0.56 for BERTScore. Using this method, a more precise clinical interpretation of the predicted report was captured from the reference report. These findings suggest that RadEval method provides a more accurate and clinically aligned framework for evaluating the medical report generation model.

1. Introduction

The automated generation of radiology reports from medical images has emerged as a significant challenge at the intersection of computer vision and clinical natural language processing (NLP) [1]–[4]. With the rapid advancement of deep learning [5], [6], particularly transformer-based architectures [7]–[9], numerous models have demonstrated promising results in generating radiology reports that are both fluent and descriptive [10], [11]. However, evaluating the clinical accuracy of these generated reports remains a critical bottleneck. It is inherently challenging due to the complex nature of medical images, the subtlety of pathological cues, and the demand for clinically sound language generation. The recent success of deep learning, particularly the advent of transformer-based vision-language models (VLMs) [12]–[15], has spurred the development of numerous systems capable of generating descriptive reports with varying levels of clinical utility [16]–[19].

Several end-to-end models, such as R2Gen [16], M2Trans [20], MedViLL [21], and GLoRIA [22], have demonstrated notable progress by learning joint visual-linguistic embeddings. Some approaches, including BioViL-T [23] and RadFormer [24], further enhance performance by incorporating self-supervised learning and hierarchical reasoning. Others integrate structured medical knowledge through ontology-guided supervision or entity-aware decoding [25], [26]. More recently, large multimodal models (LMMs) like ChatCAD+ [27] have shown promise in adapting generalist language models to medical imaging tasks. Despite these advances, the evaluation of generated reports remains a persistent bottleneck.

Traditional natural language generation (NLG) metrics such as BLEU [28], ROUGE [29], and METEOR [30] primarily assess surface-level similarity via n-gram overlap, which fails to capture clinical correctness or conceptual similarity in the medical context. Recent efforts have turned to embedding-based metrics, such as BERTScore [31] and BLEURT [32], which attempt to measure contextual similarity but still lack domain-specific alignment. To address this, several task-specific metrics have emerged. CheXpert [33] computes label agreement over 14 thoracic conditions, while RadGraph [34] evaluates structured entities and their relationships in a graph-based format. RadCliQ [35] proposes a human-in-the-loop regression model trained to mimic expert judgments, and F1CheXbert [36] combines text and label supervision to enhance robustness. Previous works have also explored ontology-guided and concept-based evaluation frameworks that measure semantic alignment at multiple levels exact match, paraphrase, and hierarchical closeness [37], [38].

However, these approaches are often limited to predefined label sets that do not generalize to rare findings or nuanced descriptors [39]. Additionally, these methods often overlook semantic phenomena such as synonymy, paraphrasing, and context-sensitive variation, leading to insensitivity to fine-grained semantic alignment and reduced robustness [40]–[42]. While medical ontologies such as RadLex [43], SNOMED CT [44], or UMLS [45], which can facilitate concept-level normalization, hierarchical reasoning, and synonym mapping they remain underutilized in current evaluation metrics. This limits their ability to capture concept equivalence across varying lexical forms (e.g., “apical pneumothorax” vs. “air in lung apex”, "cardiomegaly" vs "enlarged heart") that may be semantically equivalent in a medical context, yet completely mismatched under n-gram matching schemes. This semantic gap leads to unreliable assessments of the clinical correctness of automated radiology reports.

To address these limitations, recent work has explored entity-based and ontology-grounded evaluation methods that map report content to structured medical knowledge bases such as UMLS and RadLex. Inspired by this direction, we propose a semantic-aware evaluation framework that integrates three key components: (1) entity extraction from generated and reference reports, (2) ontology normalization via RadLex concept mapping, and (3) contextual clustering of semantically similar entities using BioBERT embeddings. Additionally, we incorporate predicted abnormality tags into the scoring mechanism to weight clinically relevant entities, ensuring that the final evaluation emphasizes findings of greater diagnostic significance.

Our method computes similarity scores at multiple levels: exact entity match, ontology ID match, and context-cluster similarity, then modulates these scores with tag confidence to reflect clinical importance. The resulting metric provides a more nuanced and robust evaluation of report quality, aligning better with human radiologist judgment and clinical relevance. To ensure practical relevance, RadEval is designed not only as a research evaluation tool but also as a component that can be seamlessly integrated into real-world clinical AI development pipelines. In this context, RadEval can serve three complementary roles: (1) as a clinical validation tool, providing ontology- and semantics-aware comparisons between generated and reference reports to assess whether models produce clinically meaningful outputs; (2) as an automated benchmarking platform, enabling fair and standardized evaluation across different models, datasets, and tasks; and (3) as a supportive module in AI deployment, where ongoing monitoring of generated reports is required to maintain quality and safety before clinical integration. By positioning RadEval within these practical workflows, we highlight its potential to bridge the gap between research evaluation and clinical applicability, offering a scalable and extensible solution

for future clinical NLP and radiology AI systems. For instance, it can function as a validation component during model training, ensuring that generated outputs align with standardized medical terminology; as an automated benchmarking platform, enabling consistent comparison of different systems across datasets; and as a monitoring tool in deployment scenarios, where ongoing quality assurance is required to maintain safe and reliable clinical performance. By situating RadEval within these practical workflows, we emphasize its potential to bridge the gap between research evaluation and real-world clinical application.

To validate our method, we conduct experiments on publicly available radiology datasets, comparing against conventional evaluation metrics and analyzing performance under different clustering and scoring settings. We also present ablation studies to assess the contribution of each module (ontology mapping, clustering, and tag weighting) to the overall metric. To address these gaps, we propose a semantic scoring framework that evaluates radiology reports by integrating medical entity extraction, RadLex-based ontology normalization, contextual clustering of entity embeddings, and clinical weighting via tag-specific importance scores. Our main contributions are:

- to perform concept-level normalization of extracted entities using RadLex IDs to enable clinically meaningful alignment beyond lexical overlap.
- to construct semantic clusters of entities based on contextual similarity using BioBERT embeddings and K-Means, enabling the method to tolerate paraphrasing and report variability.
- to introduce clinically informed weighting based on tag importance, which prioritizes alignment on abnormal or diagnostically significant findings.
- to compute matching scores at multiple levels, exact, ontology-concept, and context-cluster, and combine them with tag scores to produce a final interpretive score, improving evaluation sensitivity.

Through these contributions, we provide a robust and interpretable framework that bridges the gap between general-purpose NLG evaluation and the unique requirements of medical report assessment in medical imaging AI. The remainder of this paper is structured as follows: Section 2 describes our proposed method, including entity extraction, ontology mapping, and contextual clustering. Section 3 presents results, comparison studies, and ablation analyses. Section 4 concludes with limitations and future work in Section 5.

2. Method

2.1. General Pipeline

We evaluate our proposed method using a publicly available chest X-ray report dataset derived from the MIMIC-CXR, which contains 227,827 paired image-caption data annotated by radiologists. For this study, we focus exclusively on the radiology report text, utilizing both the original ground truth reports and their corresponding model-generated predictions. Given two radiology reports, the reference report and the predicted report, the proposed method computes a semantic similarity score by leveraging entity extraction, medical ontology alignment, and context-based clustering. The process begins by extracting predefined medical entities from both reports using rule-based matching techniques. To resolve lexical ambiguity and improve consistency, each extracted entity is mapped to a standardized concept ID using a RadLex-based ontology mapping. This mapping enables the system to treat different surface forms of the same concept as equivalent.

To further account for semantic similarity beyond surface form or ontology alignment, contextual embeddings of entities are constructed. For each entity, all sentences containing the entity across the corpus are aggregated, and sentence embeddings are computed using BioBERT. These contextual embeddings are then averaged and clustered using K-Means, allowing semantically similar entities to be grouped based on their usage in similar clinical contexts. Entity-level similarity is evaluated at three granularity levels:

- Exact Match (EM): the predicted entity and the reference entity are identical.

- Concept Match (CM): the entities differ lexically but map to the same ontology concept ID.
- Cluster Match (CL): the entities have different labels and ontology IDs, but belong to the same context cluster based on embedding similarity.

For each predicted-reference entity pair, a similarity score is assigned as (1), where is the set of entities extracted from the predicted report, and is the set of entities extracted from the reference report. To incorporate model confidence in entity prediction, each similarity score is weighted by the corresponding predicted tag confidence (tag score). This leads to a final match score per entity, formulated as (2), where the label abnormalities from the predicted report are used; these scores are aggregated across all entity pairs in the report to compute semantic-aware performance metrics. Precision, recall, and F1-score are calculated using the number of true positives (TP), false positives (FP), and false negatives (FN), where a true positive is defined as a predicted entity that matches a reference entity at any of the defined similarity levels.

$$SimilarityScore(e_p, e_r) = \begin{cases} 1.0 \; if \; e_p = e_r \\ 0.9 \; if \; ConceptID\,(e_p) = ConceptID(e_r) \\ 0.7 \; if \; Cluster(e_p) = Cluster(e_r) \\ 0.0 \; otherwise \end{cases} \quad (1)$$

$$FinalMatchScore\,(e_p, e_r) = SimilarityScore(e_p, e_r) * TagConfidence(l_p) \quad (2)$$

This approach enables a more nuanced evaluation of generated radiology reports by rewarding semantically aligned yet lexically divergent predictions while accounting for model certainty. This approach is also illustrated in Fig. 1.

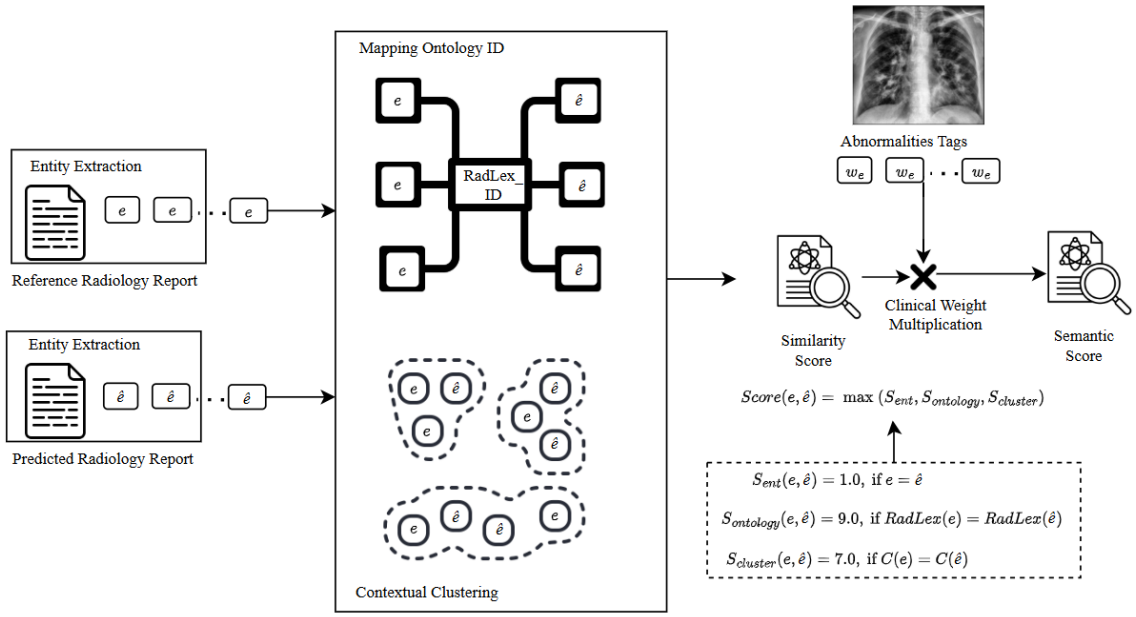

Fig. 1. Proposed model scheme

2.2. Entity extraction

To assess semantic similarity between radiology reports, entity extraction was performed on both the predicted and ground-truth reports. The aim is to identify clinically relevant terms such as anatomical structures, pathologies, and clinical findings that serve as the basis for comparison. This process uses rule-based matching via exact string matching and regular expressions. A pre-trained NER model from Hugging Face was used and tuned on the same dataset. Post-processing step is conducted to normalize the entities, including lowercasing, lemmatization, and mapping variations of expressions (e.g., "opacity" vs. "opac") to a standardized form. Additionally, boundary-aware regular expressions are applied to

eliminate partial matches and ensure entity-level granularity. This step ensures that only well-defined medical terms are extracted, providing a reliable set of entities for further semantic evaluation.

Given the diversity and variability of clinical language in radiology reports, entity extraction requires careful tuning to the target dataset. Several adjustments were made.

- **Vocabulary Curation:** The entity list was refined through an iterative review of the dataset, ensuring inclusion of frequent and clinically significant terms while avoiding overly general or ambiguous words.
- **Normalization:** Variants of terms (e.g., "effusion" vs. "pleural effusion") were standardized to a canonical form where appropriate to increase extraction consistency.
- **Ambiguity Resolution:** Terms with potential for cross-domain ambiguity were filtered or disambiguated via context rules or ontology mapping during later stages.
- **Noise Reduction:** To handle misspellings and rare typographical errors in free-text reports, lightweight preprocessing was applied (e.g., extra space removal, punctuation handling).

The output of this process is a robust entity set per report, forming the foundation for ontology alignment and clustering in the semantic evaluation pipeline.

2.3. Mapping Ontology ID

After the entities were extracted from the medical report, each was generated, and a ground truth report was created. The entities extracted were mapped using ontology ID based on the unique concept identifier RadLex, a comprehensive lexicon for radiological terms developed by the Radiological Society of North America (RSNA). RadLex provides a comprehensive lexicon of radiological terms, enabling consistent representation of clinical concepts across reports. By linking entities to their corresponding concept_id, this step serves as a normalization mechanism that aligns semantically for similar terms (e.g., "cardiomegaly" and "enlarged heart") under a shared identifier. To align extracted medical entities with standardized terminology, we map each entity to its corresponding concept in the RadLex ontology. This process enables semantic interoperability by replacing ambiguous entities with RadLex-concept identifiers.

For each extracted entity, a direct match is sought in the mapping by querying the RadLex API via BioPortal. Only entities with valid concept identifiers are retained for downstream processing, ensuring that all considered terms are grounded in medical ontology. This ontology alignment supports concept-level matching in later evaluation steps, allowing the metric to reward entities from the predictions report that are semantically correct, even if the lexical form differs from the reference report. When a direct match is not found, fallback strategies are applied to handle ambiguous or unmatched terms. Fuzzy string matching captures lexical variations, such as abbreviations or spelling differences, and semantic similarity scoring with BioBERT embeddings identifies the closest RadLex concept based on contextual meaning. If no suitable match is determined after these steps, the entity is treated as unmatched and excluded from downstream scoring.

Once RadLex IDs have been assigned to entities, they serve as canonical representations of the underlying clinical concepts. During evaluation, a predicted entity and a reference entity are considered semantically aligned if they share the same RadLex ID, regardless of the textual differences. This normalization step facilitates: 1) Ontology-aware scoring, where a partial match can still contribute to evaluation metrics if the two entities share the same concept ID; and Reduced lexical ambiguity, improving the robustness of semantic similarity evaluation across diverse reporting styles and terminologies. Concept normalization via RadLex enables the model to be evaluated not just on exact textual matches, but also on clinically meaningful equivalence, thereby increasing the medical validity of the scoring process.

2.4. Cluster Semantic Awareness

To further improve semantic alignment between predicted and reference reports, a contextual clustering mechanism that groups semantically similar entities based on their usage in context were

implemented. This step accounts for subtler similarities not captured by lexical matching or ontology normalization alone.

For each unique entity extracted from the combined set of predicted and reference reports, all sentences or report segments in which the entity appears were aggregated. These surrounding texts represent the context in which the entity is used. This context provides essential cues that help differentiate entities with ambiguous meanings or refine the understanding of a specific term in a particular clinical usage.

Formally, as formulated in Eq. (3) let e be an entity from the set of all extracted entities. We define $C_e = \{x_i | e\ x_i\}$ as the collection of report segments where e occurs. To encode the contextual semantics of each entity, we use a pre-trained biomedical language model, BioBERT, to generate contextual sentence embeddingSs. For each report segment $x_i C_e$, we compute the embedding $h_i\ R^d$ using the [CLS] token representation from the final hidden layer of BioBERT. The entity's final representation h_e is obtained by averaging all contextual embeddings in C_e as can be seen in (3).

$$\bar{h}_e = \frac{1}{|C_e|} \sum_{x_i \in C_e} BioBERT_{CLS}(x_i) \tag{3}$$

This means embedding captures the typical context in which the entity appears, serving as a robust, usage-aware representation. Once the embeddings for all entities were obtained, unsupervised K-Means clustering was used to group entities with similar contextual embeddings into semantic clusters. This step aims to capture indirect or latent semantic relations, such as "pleural thickening" and "pleural calcification" appearing in similar diagnostic narratives but having different lexical forms and RadLex IDs, denoted as (4).

$$C(e): E\ \{1,2, \dots, K\} \tag{4}$$

where K is the number of clusters, empirically chosen based on the dataset distribution. The contextual clustering mechanism enables cluster-level matching during evaluation. Even if two entities differ lexically and semantically (i.e., no exact or ontology match), a shared cluster membership implies contextual similarity. Therefore, it contributes to the evaluation score as a weaker form of semantic alignment: exact matches provide the strongest signal of correctness, ontology matches based on shared RadLex IDs offer a moderate degree of semantic alignment, and cluster-level matches, while weaker, still supply contextually relevant evidence that the predicted term resides in a similar semantic neighborhood as the ground-truth label. By incorporating contextual clustering, the evaluation framework captures nuanced and clinically meaningful entity similarity beyond surface-level matching.

2.5. Scoring Calculation

Scoring is calculated by computing precision and recall for each level of similarity, from entity-level to cluster-level matching. To quantify the semantic alignment between the predicted and ground truth reports, a similarity-based scoring function is computed that combines exact lexical matching, ontological concept equivalence, and contextual clustering, each modulated by a clinical importance weight derived from the predicted abnormality tags.

Each entity extracted from both is evaluated against the other on three levels:

- Exact Entity Match. If an entity that is noted as appearing in both and with the same surface form, it is considered an exact match with a score of 1 as defined in (5).

$$S_{ent}(e, \hat{e}) = \ 1.0\ if\ e = \hat{e} \tag{5}$$

- Ontology Match (RadLex ID). If two entities map to the same RadLex concept ID but differ lexically, the match is considered semantically equivalent via medical ontology as defined in (6).

$$S_{ontology}(e, \hat{e}) = \ 0.9\ if\ RadLex(e) = RadLex(\hat{e}) \tag{6}$$

- Contextual Cluster Match. If two entities do not match directly or ontologically but belong to the same contextual cluster, they are considered weakly aligned as defined in (7)

$$S_{cluster}(e, \hat{e}) = \ 0.7 \ if \ C(e) = C(\hat{e}) \quad (7)$$

If none of the above conditions hold, then the score is valued as 0

- Clinical Weighting with Abnormality Tags. To reflect clinical relevance, each reference entity e is weighted by a tag importance score $w_e \in [0,1]$, which is derived from the model's predicted confidence in abnormality detection. This modulates the impact of each match according to its perceived clinical importance.

The final semantic score for a pair of entities $(e, \hat{e})$ is computed as stated in (8).

$$Score(e, \hat{e}) = \max(S_{ent}, S_{ont}, S_{cluster}) \times w_e \quad (8)$$

- Report-Level Evaluation. For each predicted–reference report pair, all possible entity matches are evaluated. True positives (TP), false positives (FP), and false negatives (FN) are calculated based on the presence or absence of matches above a threshold. Precision, recall, and F1-score are then computed by (9).

$$Precision = \frac{TP}{TP+FP}, \ Recall = \frac{TP}{TP+FN}, \quad (9)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (10)$$

This results in a clinically aware, semantically grounded evaluation score that captures the meaningful overlap between reports.

3. Results and Discussion

This section describes the results and analysis of each process in the scoring level evaluation:

3.1. Extracted Entities

A total of 171 unique clinical entities were extracted from the radiology report corpus. The extraction was performed using a fine-tuned biomedical NER model to ensure alignment with the radiology domain. Following extraction, a normalization step was applied to unify synonymous expressions across reports. This final entity set served as the basis for the subsequent RadLex mapping and contextual similarity assessment.

3.2. Analysis of Ontology mapping ID process

Following the entity extraction phase, ontology-based normalization was conducted to unify semantically equivalent terms across radiology reports. This was achieved by mapping each entity to a standardized concept identifier (ID) using the RadLex ontology, accessed via the NCBO BioPortal API. The RadLex ontology, widely used in radiology informatics, provides hierarchical and semantic definitions of anatomical structures, imaging observations, and diagnostic terms. Of the 171 extracted entities, 112 were successfully aligned with unique RadLex concept IDs, covering both anatomical and observational categories.

The ontology mapping process aims to obtain lexically diverse but semantically identical terms under a common concept. For example, both *cardiomegaly* and *cardiac enlargement* were mapped to RID28764, indicating their shared clinical meaning. Similarly, *fibrosis* and *pulmonary fibrosis* were unified under RID3121. In some cases, multiple entities representing fine-grained variants or synonyms (e.g., *heart enlargement*, *enlarged heart*, *cardiomegaly*) were resolved to a single identifier, thereby reducing ambiguity and enabling robust semantic alignment. The sample of mapped data can be seen in Table 1.

Table 1. Ontology mapping example

Entity example	Normalization	RadLex ID	Medical Concept	Semantic
Right lung	right lung	RID35866	Lung	Body Part, Component Organ
Consolidation	consolidation	RID10346	Pulmonary Consolidation	Observation
Pleural fluid	pleural fluid	RID32225	Pleural Effusion	Observation
Cardiomegaly	cardiomegaly	RID28764	Heart Enlargement	Observation
Cardiac Enlargement	cardiac enlargement	RID28764	Heart Enlargement	Observation

Through this ontology mapping, lexical variability was successfully minimized, allowing for a more standardized semantic comparison between reports. By grounding extracted entities in RadLex IDs, the framework ensured that semantic comparisons between predicted and reference reports were based on clinically relevant concepts rather than surface-form string similarity.

3.3. Contextual-based clustering

In addition to ontology mapping with RadLex IDs, the extracted entities were further grouped based on contextual semantics using clustering techniques. In the process of contextual grouping, multiple clustering algorithms, K-Means, Agglomerative Clustering, and DBSCAN were applied to entity-level contextual embeddings. These embeddings were obtained using the [CLS] token representation from the BioBERT model. For each entity, multiple sentence contexts (10–20 per entity) were sampled from the dataset, and their embeddings were averaged to construct a representative semantic vector.

To determine the optimal clustering approach, quantitative and qualitative analyses were performed. The quality of the resulting clusters was quantitatively evaluated using the Silhouette Score and the Davies-Bouldin Index (DBI) for evaluating cluster cohesion and separation, as can be seen in Table 2. K-Means achieved the highest silhouette scores, indicating better intra-cluster similarity and inter-cluster separation compared to the alternatives. K-Means consistently produced more balanced clusters with interpretable clinical groupings such as anatomical locations (e.g., *apex*, *lung base*), pathological conditions (e.g., *pneumothorax*, *effusion*), and clinical descriptors (e.g., *acute*, *chronic*). Agglomerative clustering, while slightly more accurate in small data regimes, often merged dissimilar clinical concepts, thereby reducing contextual precision. Meanwhile, DBSCAN underperformed due to its sensitivity to parameter tuning and poor handling of high-dimensional sparse entity spaces.

Table 2. Evaluation of the cluster method

Clustering Method	Silhouette Score↑	Davis-Bouldin Index ↓
K-Means (k=3)	0.471	0.82
Agglomerative (Ward)	0.398	1.07
DBSCAN (eps=0.6)	0.211	1.59

Our contextual clustering was visualized using BioBERT-derived entity embeddings with t-SNE to analyze the semantic quality. Fig. 2 illustrates the resulting 2D projections, with each point representing an entity, and colors indicating cluster assignments from different algorithms. Notable cluster groupings included:

- Anatomical Structures with light blue color dot: “apex”, “diaphragm”, “lung base”
- Pathological Conditions with dark blue color dot: “pneumothorax”, “effusion”, “consolidation”
- Clinical Descriptors with brown color dot: “acute”, “moderate”, “diffuse”

As shown in Fig. 2, K-Means produced compact, semantically coherent clusters that align with clinical groupings and are tight and interpretable. In contrast, Agglomerative Clustering occasionally merged anatomically and pathologically unrelated terms, forming larger, less precise clusters, mixing

anatomical terms with descriptors likely due to chaining effects. DBSCAN, though parameter-free in clustering count, struggled with the sparse embedding space and failed to assign valid clusters, creating many singleton points and failing to identify meaningful groups. This visualization shows the semantic cohesion achieved by the proposed approach and highlights the limitations of traditional clustering algorithms in biomedical embedding spaces.

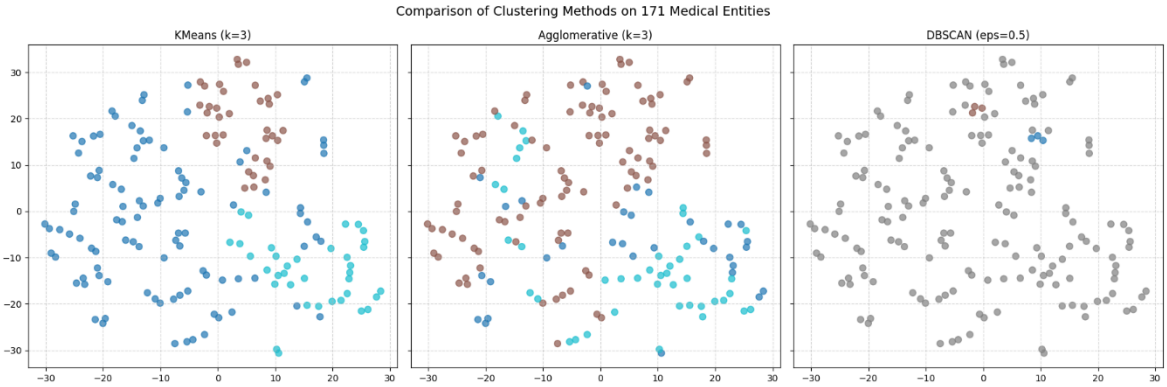

Fig. 2. Visualization of 2D data in different cluster methods

3.4. Result analysis

This section describes the analysis of the scoring process. Table 3 provides the contribution of each component in RadEval using the F1-score. It was observed that including cluster-level contextual similarity improved the F1-score by +0.19, confirming its ability to capture semantic matches beyond direct ontology mappings. When combined with abnormality tag weighting, further gains were achieved, with the final evaluation better emphasizing clinically relevant findings while reducing the influence of noise or ambiguous descriptors. The results demonstrate that incorporating ontology-level and contextual clustering significantly improves semantic alignment. Furthermore, weighting the scores by predicted tag scores (clinical confidence) leads to the highest improvement, reflecting better prioritization of clinically important matches.

Table 3. Evaluation of each component to model performance

Evaluation Strategy	Precision	Recall	F1-Score
Exact match only	0.44	0.39	0.41
+ RadLex concept matching	0.55	0.52	0.53
+ Cluster-level contextual matching	0.63	0.58	0.60
+ Tag-weighted final scoring (ours)	0.67	0.62	0.64

3.5. Comparison with existing method

When compared to conventional string-based metrics such as BLEU and ROUGE, the proposed method exhibited a stronger correlation, as clinically observed. Traditional n-gram-based metrics often failed to reward clinically valid paraphrases or synonyms, whereas the proposed semantic score captured latent similarities via RadLex mapping and contextual grouping. For example, in one case, a predicted report that referenced "pleural fluid" instead of "pleural effusion" received a low BLEU score due to n-gram mismatch, but was correctly matched under the proposed method owing to the shared RadLex ID.

To comprehensively evaluate the clinical validity and semantic alignment of generated reports, we conducted a comparative analysis using multiple existing evaluation methods alongside our proposed framework. Metrics examined included BLEU, ROUGE, METEOR, BERTScore, CheXpert label matching, and RadGraph graph-based evaluation. From one of the report data "*There is cardiomegaly with evidence of pulmonary vascular congestion. No pleural effusion is seen.*" that predicted as "*The heart is enlarged and signs of pulmonary edema are present. No fluid in the pleural space.*" with tags abnormalities "*cardiomegaly*", "*pulmonary edema*", and "*pleural effusion*". The comparison can be seen in Table 4.

Table 4. Comparison of several evaluation metrics on example report

Evaluation Method	Score	Strengths	Limitations
BLEU-4	0.32	Captures surface-level n-gram overlap	Penalizes paraphrase; not domain-aware
ROUGE-L	0.41	Captures the longest common subsequence	Sensitive to word order and phrasing
METEOR	0.52	Handles synonymy and partial matches	Still string-based; lacks clinical grounding
BERTScore (F1)	0.84	Embedding-based semantic matching	Ignores ontology and clinical importance
CheXpert Labeler	3/3	Accurately extracts predefined clinical labels	Limited to 14-class label space
RadGraph F1	0.77	Entity-relation-aware; strong graph grounding	Limited relation types, no soft paraphrase matching
Proposed Semantic Score	0.86	Ontology-aware, context-aware, clinically weighted	Requires RadLex mapping + clustering

Several example cases were analyzed to illustrate these improvements further:

- In one case, *"enlarged heart"* (prediction) was mapped to the same RadLex ID as *"cardiomegaly"* (reference), yielding a high ontology match.
- In another, the predicted *"lower lobe opacity"* was not an exact match but fell into the same cluster as *"lung base infiltration"*, producing a non-zero score that BLEU would penalize entirely.
- Common errors include ambiguous terms (*"increased markings"*) that were not well-clustered, indicating that domain-specific fine-tuning of embeddings could improve future contextual modeling.

These examples further validate the semantic sensitivity of the proposed approach and its alignment with clinical interpretability. The ablation study was also done to confirm that each component, ontology mapping, contextual clustering, and tag-based weighting, contributes meaningfully to the overall improvement in semantic evaluation. In particular, the removal of tag-based weighting led to a substantial reduction in performance, highlighting the importance of incorporating clinical prioritization into the scoring process. The comparison is shown in Table 5.

Table 5. Comparison based on the semantic sensitivity of different methods

Metric	Type	Captures Semantic Similarity	Captures Clinical Importance	Ontology Awareness	Interpretability
BLEU	N-gram Overlap	✗	✗	✗	✗
ROUGE-L	Longest Common Subseq.	✗	✗	✗	✗
METEOR	N-gram + Synonyms	✓ (partially)	✗	✗	✗
BERTScore	Embedding Similarity	✓	✗	✗	✗
CheXpert F1	Disease Label Matching	✓	✓	✗	✓
RadGraph F1	Entity-Relation Match	✓	✓	✓ (structure)	✓
Ours	**Tag-aware Semantic Match**	✓	**✓ (via tag weights)**	**✓ (via RadLex)**	**✓ (high)**

A comparison of the metrics that indicate the semantic score is also provided in Table 6, which shows the score values for each metric.

Table 6. Comparison with other evaluation methods

Metric Type	Metric	Score
Text Similarity	BLEU-4	0.12
	ROUGE-L	0.28
Embedding Similarity	BERTScore (F1)	0.56
Semantic Evaluation	Proposed Semantic Score (F1)	0.64
Ablation Study	w/o Tag Weighting (F1)	0.60
	w/o Clustering (F1)	0.53
	w/o Ontology Mapping (F1)	0.41

We also compare using the existing method with our evaluation framework. For each technique, compare each level of semantic scoring that can be seen in Table 7.

Table 7. Evaluation with other generation methods

Model	Evaluation	F1-Score	Precision	Recall
R2Gen	Exact Match	0.34	0.35	0.33
	RadLex Concept Match	0.51	0.52	0.50
	Cluster-Level Contextual Match	0.61	0.62	0.60
R2GenCMN	Exact Match	0.47	0.48	0.46
	RadLex Concept Match	0.66	0.67	0.65
	Cluster-Level Contextual Match	0.78	0.79	0.77
Proposed Model	**Exact Match**	**0.41**	**0.42**	**0.40**
	RadLex Concept Match	**0.55**	**0.56**	**0.54**
	Cluster-Level Contextual Match	**0.63**	**0.65**	**0.61**

The results demonstrate a consistent trend across all models: evaluation scores (F1-Score, Precision, and Recall) significantly improve as the evaluation method moves from a strict Exact Match to more semantically aware levels such as RadLex Concept Match and Cluster-Level Contextual Match. This highlights the limitations of traditional, surface-level metrics and underscores the need for a semantic evaluation framework. The evaluation also demonstrates distinct performances; all models achieve their lowest scores under Exact Match, reflecting the strict lexical constraints of word-level overlap. R2GenCMN achieves the best performance in this setting thanks to its memory network, which enables better retention of specific phrasing. In comparison, the Proposed Model shows moderate Exact Match performance, outperforming R2Gen but falling below R2GenCMN.

However, when shifting from strict lexical matching to ontology-aware evaluation (RadLex Concept Match), all models show notable score increases, indicating that many clinically correct findings are expressed using different phrasing or synonyms in the generated reports. Under this setting, the Proposed Model achieves a meaningful improvement over R2Gen (F1: 0.55 vs. 0.51), suggesting stronger capability in capturing medically equivalent concepts even when the lexical form differs.

The performance differences become more pronounced under Cluster-Level Contextual Match, which incorporates contextual semantic similarity. In this evaluation, the Proposed Model increases its F1-score to 0.63, again surpassing R2Gen (0.61) and demonstrating that its generated content aligns better with the clinical meaning of the reference reports even when surface forms differ. Although R2GenCMN remains the strongest model, the Proposed Model exhibits competitive contextual performance and narrows the gap, indicating that integrating image enhancement, transfer learning, and multi-label abnormality prediction contributes positively to semantic alignment.

Despite its promising results, several limitations remain. Based on the clustering granularity, the current method uses a fixed number of clusters, which may not optimally reflect the proper semantic granularity of medical concepts. Dynamic or hierarchical clustering approaches may provide better grouping. Although tag confidence is incorporated, it is uniformly applied. In practice, clinical impact varies significantly across conditions (e.g., pneumothorax vs. mild scoliosis), suggesting a need for

domain-specific clinical weighting schemes. Furthermore, the current implementation of RadEval relies primarily on English-based resources (RadLex ontology and BioBERT embeddings). While effective for English-language radiology reports, this poses a limitation in multilingual settings. However, the framework itself is extensible: by substituting RadLex with multilingual ontologies (e.g., SNOMED CT, UMLS) and adopting multilingual biomedical embeddings (e.g., multilingual BioBERT, XLM-R), RadEval can be adapted to diverse linguistic contexts, which we identify as a promising avenue for future research.

A further limitation concerns the reliability of ontology mapping. While most entities are successfully aligned to RadLex concepts, some terms cannot be matched directly. In these cases, fallback strategies such as fuzzy string matching and BioBERT embedding similarity are used; however, entities that remain unmatched are excluded from evaluation. This exclusion may bias the metric toward more common or well-defined concepts, underestimating performance on rarer or less standardized terms. Future work will explore integrating larger ontologies such as UMLS or SNOMED CT, as well as improving disambiguation techniques, to minimise the impact of unmatched terms. Additionally, the current implementation primarily relies on English-language resources (RadLex and BioBERT). While effective for English radiology reports, this restricts multilingual applicability. We note that the framework is extensible and can be adapted to other languages by incorporating multilingual ontologies and biomedical embeddings (e.g., multilingual BioBERT and XLM-R).

Another significant limitation is the lack of direct alignment with expert radiologist evaluations in the current study. Our dataset did not contain multiple expert annotations to allow inter-rater agreement analysis. While RadEval demonstrated strong semantic consistency with reference reports, validation against expert clinical judgment remains essential. Future work will include incorporating expert radiologist scoring and inter-rater agreement studies to more conclusively establish RadEval's reliability and clinical relevance. Given these limitations, several future directions remain to be explored, such as contrastive learning techniques to generate semantically consistent clusters based on report similarity. Additionally, a detailed correlation analysis between the proposed metric and expert radiologist assessments would further validate the clinical relevance of this semantic evaluation approach.

4. Conclusion

This study presents a semantic evaluation framework for medical report generation that integrates medical entities, ontologies, and contextual embeddings to assess clinical relevance. By leveraging RadLex for ontology mapping, BioBERT for context-sensitive representation, and unsupervised clustering to group semantically similar entities, the proposed method showed a more meaningful measure of clinical accuracy. The inclusion of tag confidence weighting further enhances the evaluation's sensitivity to clinically relevant abnormalities, enabling a graded assessment of model predictions. Experiments demonstrate that this framework achieves higher correlation with human expert judgment than traditional metrics such as BLEU or ROUGE, particularly in cases involving paraphrasing, synonymy, or latent semantic overlap. Our study validates the necessity of each component: ontology matching, contextual clustering, and tag-based weighting all contribute significantly to overall scoring performance. This layered, interpretable evaluation pipeline sets a new direction for assessing clinical report generation systems, focusing on semantic and clinical validity rather than surface-level similarity. Importantly, although this work focused on chest X-ray reports, the underlying framework is domain-agnostic and can be adapted to other imaging modalities, such as CT or MRI, or to broader clinical report domains by substituting the ontology and embedding resources with domain-appropriate counterparts.

Acknowledgment

This research was funded by the Ministry of Education and Culture of the Republic of Indonesia with a Pendidikan Magister menuju Doktor untuk Sarjana Unggul (PMDSU) scholarship program under the Penelitian Disertasi Doktor (PDD) scheme.

Declarations

Author contribution. The contribution or credit of the author must be stated in this section.
Tsaniya Hilya: Conceptualisation, Methodology, Investigation, Data Curation, Writing – Original Draft, Formal Analysis, Visualisation, Writing – Review & Editing.
Fatichah Chastine: Supervision, Methodology, Investigation, Formal Analysis, Writing – Review & Editing, Funding Acquisition.
Suciati Nanik: Supervision, Methodology, Validation, Writing – Review & Editing.
Funding statement. The funding agency should be written in full, followed by the grant number in square brackets and the year.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] G. Reale-Nosei, E. Amador-Domínguez, and E. Serrano, "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," *Med. Image Anal.*, vol. 97, p. 103264, Oct. 2024, doi: 10.1016/j.media.2024.103264.

[2] U. Berger, G. Stanovsky, O. Abend, and L. Frermann, "Surveying the Landscape of Image Captioning Evaluation: A Comprehensive Taxonomy, Trends and Metrics Analysis," *arxiv Artif. Intell.*, pp. 1–44, Mar. 2025. [Online]. Available at: https://share.google/cl5FrpQPS8mhk88ck.

[3] D. Sharma, C. Dhiman, and D. Kumar, "Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey," *Expert Syst. Appl.*, vol. 221, p. 119773, Jul. 2023, doi: 10.1016/j.eswa.2023.119773.

[4] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023, doi: 10.1038/s41586-023-05881-4.

[5] K. Zhang, P. Li, and J. Wang, "A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions," *Remote Sens.*, vol. 16, no. 21, p. 4113, Nov. 2024, doi: 10.3390/rs16214113.

[6] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artif. Intell. Med.*, vol. 106, p. 101878, Jun. 2020, doi: 10.1016/j.artmed.2020.101878.

[7] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[8] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, p. 102802, Aug. 2023, doi: 10.1016/j.media.2023.102802.

[9] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021, pp. 1–21, [Online]. Available at: https://openreview.net/pdf?id=YicbFdNTTy.

[10] H. Tsaniya, C. Fatichah, and N. Suciati, "Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement," *IEEE Access*, vol. 12, pp. 25429–25442, 2024, doi: 10.1109/ACCESS.2024.3364373.

[11] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation," 2021, pp. 72–82, doi: 10.1007/978-3-030-87199-4_7.

[12] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions in Vision Transformers for efficient training on small datasets," *Neurocomputing*, vol. 617, p. 128998, Feb. 2025, doi: 10.1016/j.neucom.2024.128998.

[13] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*, 2021, pp. 5583–5594, [Online]. Available at: https://share.google/nup9TStriM3ctqEN9.

[14] H. Tan and M. Bansal, "LXMert: Learning cross-modality encoder representations from transformers," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 5100–5111, 2019, doi: 10.18653/v1/d19-1514.

[15] M. M. Mohsan, M. U. Akram, G. Rasool, N. S. Alghamdi, M. A. A. Baqai, and M. Abbas, "Vision Transformer and Language Model Based Radiology Report Generation," *IEEE Access*, vol. 11, pp. 1814–1824, 2023, doi: 10.1109/ACCESS.2022.3232719.

[16] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating Radiology Reports via Memory-driven Transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1439–1449, doi: 10.18653/v1/2020.emnlp-main.112.

[17] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology Report Generation with frozen LLMs," *Meta-Radiology*, vol. 1, no. 3, p. 100033, Nov. 2023, doi: 10.1016/j.metrad.2023.100033.

[18] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 21315–21326, doi: 10.1109/ICCV51070.2023.01954.

[19] P. Singh and S. Singh, "ChestX-Transcribe: a multimodal transformer for automated radiology report generation from chest x-rays," *Front. Digit. Heal.*, vol. 7, Jan. 2025, doi: 10.3389/fdgth.2025.1535168.

[20] Z. Ni *et al.*, "M2Trans: Multi-Modal Regularized Coarse-to-Fine Transformer for Ultrasound Image Super-Resolution," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 5, pp. 3112–3123, May 2025, doi: 10.1109/JBHI.2024.3454068.

[21] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 12, pp. 6070–6080, Dec. 2022, doi: 10.1109/JBHI.2022.3207502.

[22] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3922–3931, doi: 10.1109/ICCV48922.2021.00391.

[23] S. Bannur *et al.*, "Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 15016–15027, 2023, doi: 10.1109/CVPR52729.2023.01442.

[24] S. Basu, M. Gupta, P. Rana, P. Gupta, and C. Arora, "RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection," *Med. Image Anal.*, vol. 83, p. 102676, Jan. 2023, doi: 10.1016/j.media.2022.102676.

[25] K. Roy, T. Garg, and V. Palit, "Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, Jun. 2023, pp. 234–236, doi: 10.1109/CAI54212.2023.00108.

[26] G. Wiher, C. Meister, and R. Cotterell, "On Decoding Strategies for Neural Text Generators," *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 997–1012, Sep. 2022, doi: 10.1162/tacl_a_00502.

[27] Z. Zhao *et al.*, "ChatCAD+: Toward a Universal and Reliable Interactive CAD Using LLMs," *IEEE Trans. Med. Imaging*, vol. 43, no. 11, pp. 3755–3766, Nov. 2024, doi: 10.1109/TMI.2024.3398350.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, p. 311, doi: 10.3115/1073083.1073135.

[29] C.-Y. Lin, "{ROUGE}: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Jul. 2004, pp. 74–81, [Online]. Available at: https://aclanthology.org/W04-1013/.

[30] A. Lavie and A. Agarwal, "Meteor," in *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 2007, pp. 228–231, doi: 10.3115/1626355.1626389.

[31] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations*, 2020, pp. 1–43, [Online]. Available at: https://openreview.net/pdf?id=SkeHuCVFDr.

[32] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7881–7892, doi: 10.18653/v1/2020.acl-main.704.

[33] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 590–597, Jul. 2019, doi: 10.1609/aaai.v33i01.3301590.

[34] S. Jain *et al.*, "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports," in *Advances in Neural Information Processing Systems*, 2021, no. NeurIPS, pp. 1–12, [Online]. Available at: https://datasets-benchmarksproceedings.neurips.cc/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

[35] F. Yu *et al.*, "Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, 2023, doi: 10.1016/j.patter.2023.100802.

[36] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1500–1519, 2020, doi: 10.18653/v1/2020.emnlp-main.117.

[37] V. N. Garla and C. Brandt, "Semantic similarity in the biomedical domain: an evaluation across knowledge sources," *BMC Bioinformatics*, vol. 13, no. 1, p. 261, Dec. 2012, doi: 10.1186/1471-2105-13-261.

[38] F. Remy, K. Demuynck, and T. Demeester, "BioLORD: Learning Ontological Representations from Definitions for Biomedical Concepts and their Textual Descriptions," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 1454–1465, doi: 10.18653/v1/2022.findings-emnlp.104.

[39] M. Gao, X. Hu, X. Yin, J. Ruan, X. Pu, and X. Wan, "LLM-based NLG Evaluation: Current Status and Challenges," *Comput. Linguist.*, vol. 51, no. 2, pp. 661–687, Jun. 2025, doi: 10.1162/coli_a_00561.

[40] D. Deutsch, R. Dror, and D. Roth, "On the Limitations of Reference-Free Evaluations of Generated Text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10960–10977, doi: 10.18653/v1/2022.emnlp-main.753.

[41] S. Min *et al.*, "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[42] D. Khashabi *et al.*, "GENIE Toward Reproducible and Standardized Human Evaluation for Text Generation," *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022*, pp. 11444–11458, 2022, doi: 10.18653/v1/2022.emnlp-main.787.

[43] "RadLex Ontology," *America, Radiological Society of North*, 2024. [Online]. Available at: https://share.google/e3HhnqQZymJCMSROV.

[44] SNOMED International, "SNOMED CT: The Global Clinical Terminology," *National Library of Medicine*, 2024. [Online]. Available at: https://www.nlm.nih.gov/healthit/snomedct/international.html.

[45] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D267--D270, 2004, doi: 10.1093/nar/gkh061.