GEMA-Score: Granular Explainable Multi-Agent Scoring Framework for Radiology Report Evaluation

Zhenxuan Zhang[1*], KinHei Lee[1*], Peiyuan Jing[1], Weihang Deng[1], Huichi Zhou[1], Zihao Jin[1], Jiahao Huang[1], Zhifan Gao[2], Dominic C. Marshall[3], Yingying Fang[1†], Guang Yang[1†]

[1]Department of Bioengineering, Imperial College London, UK
[2]School of Biomedical Engineering, Sun Yat-sen University, China
[3]Department of Surgery & Cancer, Imperial College London, UK
{z.zhenxuan24, k.lee24, peiyuan.jing22, weihang.deng24, h.zhou24, z.jin23, j.huang21, g.yang}@imperial.ac.uk
gaozhifan@gmail.com, dominic.marshall12@imperial.ac.uk, y.fang@imperial.ac.uk

Abstract

Automatic medical report generation has the potential to support clinical diagnosis, reduce the workload of radiologists, and demonstrate potential for enhancing diagnostic consistency. However, current evaluation metrics often fail to reflect the clinical reliability of generated reports. Overlap-based methods overlook fine-grained details (e.g., location, severity), diagnostic metrics are constrained by fixed vocabularies. Some diagnostic metrics are limited by fixed vocabularies or templates, reducing their ability to capture diverse clinical expressions. LLM-based metrics lack interpretable reasoning, limiting trust in clinical settings. Therefore, we propose a Granular Explainable Multi-Agent Score (GEMA-Score) in this paper, which conducts both objective quantification and subjective evaluation through a large language model-based multi-agent workflow. Our GEMA-Score parses structured reports and employs stable calculations through interactive exchanges of information among agents to assess disease diagnosis, location, severity, and uncertainty. Additionally, an LLM-based scoring agent evaluates completeness, readability, and clinical terminology while providing explanatory feedback. Extensive experiments show that GEMA-Score achieves the highest correlation with human experts on public datasets (Kendall = 0.69 on ReXVal; 0.45 on RadEvalX), demonstrating improved clinical scoring reliability.

Code — https://github.com/Zhenxuan-Zhang/GEMA_score

Introduction

Automatic medical report generation (AMRG) can support clinical diagnosis and reduce radiologists' workload, and many studies have explored high-quality report generation (Chen et al. 2020, 2024; Wu et al. 2023). As AMRG models proliferate, reliable evaluation is essential to ensure accuracy and prevent clinical risk. However, traditional metrics often fail to reflect clinical practicality, since report quality relies on comprehensive and granular expression (Banerjee and Lavie 2005; Lin 2004; Papineni et al. 2002). Chest X-ray reports must specify lesion type, location, severity, and uncertainty; missing such details may lead to misdiagnosis (Yu et al. 2022). Clear and logical language is also necessary to avoid misunderstanding. Therefore, a fine-grained and explainable evaluation metric is needed.

Existing evaluation metrics can be broadly grouped into five categories (Fig. 1): overlap-based, BERT-based, NER-F1-based, diagnostic-based, and LLM-based methods (Papineni et al. 2002; Lin 2004; Zhang et al. 2019; Jain et al. 2021; Smit et al. 2020; Wu et al. 2023; Ostmeier et al. 2024; Xie et al. 2024; Zhao et al. 2024). These metrics differ in linguistic granularity, clinical relevance, and interpretability. Overlap-based metrics (e.g., BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004)) rely on exact n-gram matches but fail to handle synonyms, minor phrasing differences, or contradictions (Boag et al. 2021; Yu et al. 2022). BERT-based metrics (e.g., BERTScore (Zhang et al. 2019)) use contextual embeddings to assess semantic similarity, but often overestimate scores and miss clinical correctness. NER-F1-based metrics (e.g., RadGraphF1 (Jain et al. 2021)) focus on medical entities but ignore synonymy and entity correctness. Diagnostic-based metrics (e.g., CheXbert (Smit et al. 2020) and Radbert (Yan et al. 2022)) classify predefined conditions, but are constrained by label scope and annotation costs (Wu et al. 2023). LLM-based metrics (e.g., GREEN (Ostmeier et al. 2024), DocLens (Xie et al. 2024), RaTE (Zhao et al. 2024)) aim to offer holistic evaluation but often lack interpretability and condition-level attribution (Gu et al. 2024; Pal, Umapathi, and Sankarasubbu 2023). Given the limitations of existing metrics, RadCliQ (Yu et al. 2022) integrates human evaluation criteria and correlates well with expert scores. Yet it remains coarse-grained and annotation-heavy.

Current evaluation faces several fundamental challenges that undermine its reliability and clinical utility. First, most metrics fail to account for clinically relevant variations in expression, such as synonyms (e.g., "opacity" vs. "infiltrate"), uncertainty modifiers ("likely", "possible"), or severity descriptors ("mild", "severe") (Boag et al. 2021; Yu et al. 2022). These nuances are critical in clinical interpretation, and their omission or misclassification can lead to inaccurate scoring and overlook important differences in diagnos-

[*]These authors contributed equally.
[†]Corresponding authors.

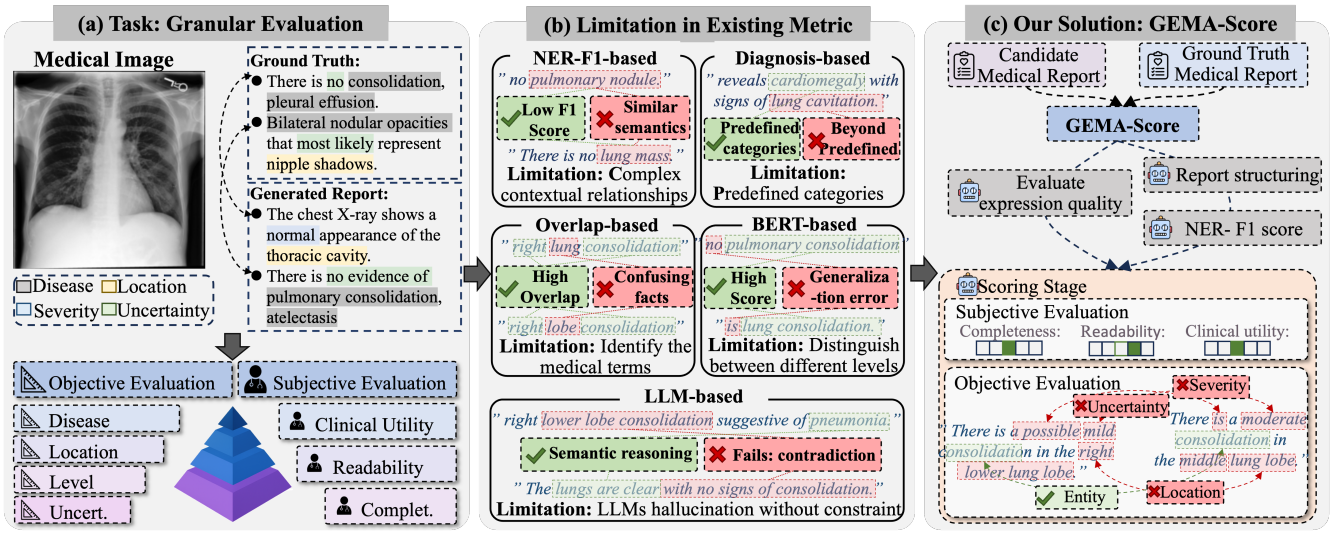

Figure 1: Motivation of our GEMA-Score. (a) The task of evaluating generated medical reports using objective and subjective metrics. (b) Limitations of existing evaluation metrics, including NER-F1, overlap-based, and BERT-based methods. (c) The proposed GEMA-Score provides a comprehensive assessment of generated reports.

tic meaning. Second, existing methods often provide aggregate scores without clear attribution, making it difficult to interpret results or trace errors back to specific report elements (Ostmeier et al. 2024; Xie et al. 2024; Zhao et al. 2024). This lack of transparency limits their usefulness in model debugging, error analysis, or human-in-the-loop validation. Third, many approaches rely either on fixed diagnostic labels, which constrain generalizability and adaptation to new conditions, or on single-step LLM-based judgments (Ostmeier et al. 2024). As a result, current evaluation pipelines struggle to balance clinical accuracy, interpretability, and scalability. It restricts the deployment of safe and trustworthy medical report generation.

To address the limitations of existing evaluation methods and improve the explainability, clinical fidelity, and transparency of report assessment, we propose the Granular Explainable Multi-Agent Score (GEMA-Score) (Fig. 2). Rather than producing a single opaque score, GEMA-Score decomposes the evaluation process into four specialized agents, each targeting a distinct aspect of report quality. This modular design directly tackles challenges by enabling fine-grained semantic understanding, interpretable error attribution, and flexible multi-criteria evaluation: (a) Entity Extraction Agent identifies key clinical findings (e.g., disease, location) from both generated and reference reports, supporting semantic alignment and synonym handling. (b) Objective Clinical Accuracy Agent computes F1 scores across four dimensions (disease, location, severity, and uncertainty) to capture diagnostic nuances. (c) Subjective Expressiveness Evaluation Agent provides a human-aligned assessment based on completeness, readability, and clinical utility, addressing aspects beyond factual correctness. (d) Score Evaluation Agent integrates the objective and subjective results to produce a comprehensive and interpretable final score. All agents operate automatically on input-output report pairs, enabling structured and transparent evaluation. GEMA-Score shows strong alignment with expert ratings (Kendall's $\tau = 0.69$ on ReXVal and 0.45 on RadEvalX), and improves the reliability and clinical applicability of report assessment in AMRG models. Our contributions are summarized as follows:

- We construct a granular explainable multi-agent score system. It combines objective quantification and subjective evaluation.
- GEMA-Score generates detailed explanatory feedback to improve the verification and reliability of the report evaluation.
- The experimental results show that GEMA-Score is highly consistent with human expert assessment and verify its clinical application potential.
- We further validate the generalizability of GEMA-Score on CT report data, demonstrating its robustness across imaging modalities.

Related Work

Multi-agent System. Multi-agent systems improve robustness and interpretability by assigning tasks to specialized agents (Li et al. 2024; Wang et al. 2025; Estornell and Liu 2024). This has enabled dialogue coordination, tool use, and task decomposition, with agents collaborating or critiquing to enhance reliability (Estornell and Liu 2024). For instance, AutoGen (Wu et al. 2024) uses planner, coder, and debugger agents for complex tasks, while MAD (Liang et al. 2023) improves factuality through adversarial debate. These approaches show that structured multi-agent reasoning can outperform single-agent pipelines (Wang et al. 2025). In clinical NLP, such collaboration holds promise for report evaluation, where distinct agents can focus on different cri-

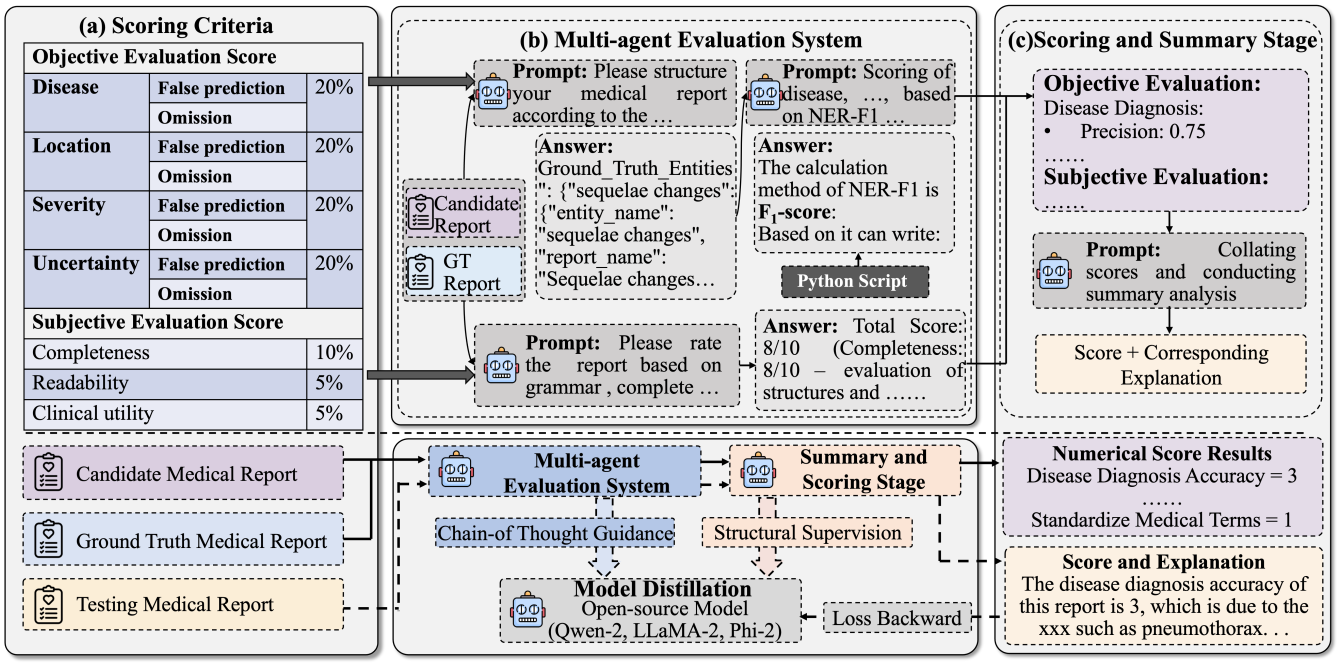

Figure 2: Workflow of our GEMA-Score. (a) The framework evaluates medical reports based on objective and subjective scoring criteria. (b) The multi-agent evaluation system assesses the candidate report against ground truth using structured prompts and automated scoring scripts. (c) The scoring and summary stage aggregates evaluation results. It provides numerical scores and detailed explanations for disease diagnosis, grammar, and terminology standardization.

teria (e.g., clinical accuracy vs. language fluency), leading to more robust and interpretable assessments.

Medical Report Evaluation Metric. Various automatic metrics have been proposed to evaluate the generated medical reports. Early metrics like BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005) rely on surface-level n-gram overlap and struggle with lexical variation, especially in clinical contexts. Semantic similarity metrics such as BERTScore (Zhang et al. 2019), MOVERScore (Zhao et al. 2019), and BLUERT (Jiang et al. 2022) leverage contextual embeddings to capture meaning, but often conflate fluency with factual accuracy and inflate scores for redundant or vague content. Structure-aware metrics such as RadGraphF1 (Jain et al. 2021) offer greater clinical relevance by extracting clinical entities and relations; however, they depend on accurate entity extraction and may miss subtle semantic errors. More recently, LLM-based evaluators aim to approximate expert judgment. GREEN (Ostmeier et al. 2024) uses free-form LLM rationales; RaTEScore (Zhao et al. 2024) and DocLens (Xie et al. 2024) incorporate aspect-specific prompts; and CheXagent (Chen et al. 2024) applies condition-focused LLM agents. Although promising, these methods still face challenges in explainable and fine-grained attribution.

Method

Problem Definition and Scoring Framework Automatic medical report generation has the potential to improve the traditional cumbersome medical process. However, a comprehensive evaluation of these report generation models remains challenging due to the dual nature of clinical assessment. On the one hand, an effective generated report must maintain objective clinical accuracy, ensuring that diseases and findings are correctly characterized. On the other hand, the report should exhibit high linguistic quality, aligning with expert radiologists' reporting styles and terminology preferences. To address these challenges, we formalize a composite scoring task with two-fold objectives. Objective evaluation S_{obj} quantifies accuracy in disease characterization based on four clinical granularities (disease entity, location, severity, uncertainty). Subjective evaluation S_{sub} assesses the completeness, readability, and clinical usability of the diagnostic report. The overall score $S_{overall}$ can be defined as:

$$S_{overall} = \sum_{i=1}^{K} w^{(i)} * S_{obj}^{(i)} + \sum_{i=1}^{I} w^{(i)} * S_{sub}^{(i)}, \quad (1)$$

where K and I represent categories defined for objective and subjective evaluations. This framework integrates both clinical precision and linguistic quality. It provides a holistic evaluation strategy for medical AI systems and bridges the gap between automated assessment and real-world radiological reporting standards.

Entity Extraction Agent. Given a generated report $\mathcal{R}$, the entity extraction agent segments it into fine-grained entities, including *disease*, *location*, *severity*, and *uncertainty*. The

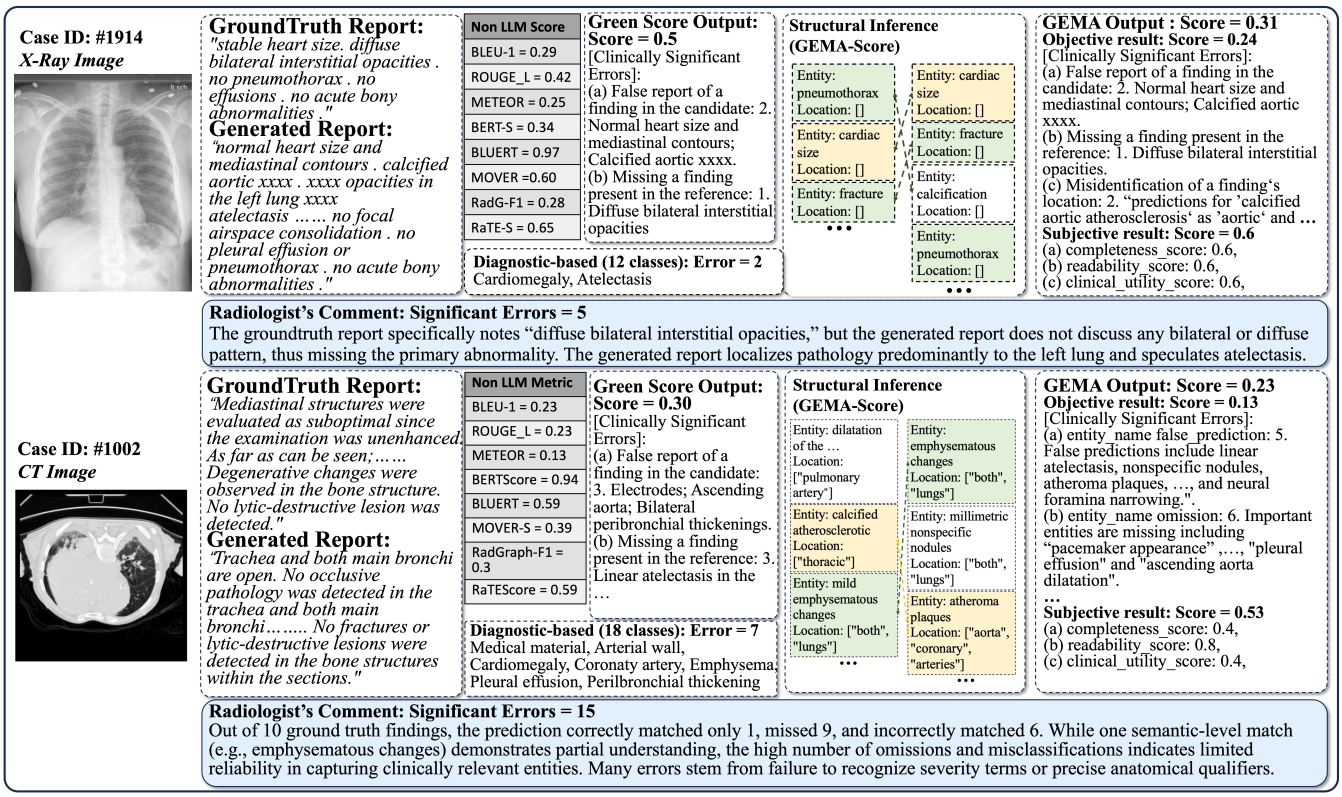

Figure 3: A multi-modal case study involving X-ray and CT images, comparing ground truth and generated reports using radiologist feedback, NLP metrics, Green-Score, and the stepwise GEMA-Score assessing clinical and linguistic quality.

structured output is represented as:

$$E = \left\{ \left(E_d^{(i)}, E_l^{(i)}, E_s^{(i)}, E_u^{(i)} \right) \right\}_{i=1}^{N}, \tag{2}$$

where $E_d^{(i)}$ denotes the disease entity, $E_l^{(i)}$ represents the corresponding anatomical location, $E_s^{(i)}$ indicates severity, and $E_u^{(i)}$ captures uncertainty descriptors. This structured representation enables granular evaluation of radiology reports.

Objective Clinical Accuracy Agent To quantitatively assess the factual consistency between the generated report $\hat{x}$ and the reference report x, we define a structured matching protocol over clinically significant entity types. Specifically, each report is decomposed into entity tuples spanning four semantic dimensions: disease, location, severity, and uncertainty. These entity sets are denoted as $E(x)$ and $E(\hat{x})$.

For each dimension, the agent performs type-specific comparisons between matched entity sets. The similarity score from reference to generation is defined as:

$$S(x, \hat{x}) = \frac{1}{|E(\hat{x})|} \sum_{e_j \in E(\hat{x})} \mathbb{I}\left(\exists e_i \in E(x) : \text{match}(e_j, e_i)\right) \tag{3}$$

where $\mathbb{I}$ is a binary condition indicator function. A symmetric comparison $S(\hat{x}, x)$ is also computed.

To reflect both precision and recall across entity types, we define the objective clinical accuracy score as the harmonic mean:

$$\mathrm{S}_{\text{obj}} = \begin{cases} 0, & \text{if } S(x,\hat{x}) + S(\hat{x},x) = 0 \\ \frac{2 \cdot S(x,\hat{x}) \cdot S(\hat{x},x)}{S(x,\hat{x}) + S(\hat{x},x)}, & \text{otherwise} \end{cases} \tag{4}$$

Each sub-dimension contributes independently and may optionally be weighted to reflect clinical importance.

Subjective Expressiveness Evaluation Agent To evaluate the linguistic quality of the generated report, we introduce a subjective expressiveness score based on fluency, grammar, and medical terminology usage. Let $\mathcal{E} = \{\text{fluency}, \text{grammar}, \text{terminology}\}$ denote the evaluated dimensions. For each aspect $a \in \mathcal{E}$, the agent identifies binary error indicators $\text{err}_a^{(k)}$ over observed issues.

The final subjective score is computed as a weighted sum:

$$\mathrm{S}_{\text{sub}} = \sum_{a \in \mathcal{E}} w_a \cdot \max\left(0, 1 - \lambda \cdot \sum \text{err}_a^{(k)}\right), \tag{5}$$

where w_a is the agent-defined importance weight and λ is penalty strength and is set to 0.05. To ensure interpretability, S_{sub} is rounded to the nearest value in the discrete set $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

Score Evaluation Agent The final GEMA-Score aggregates factual and linguistic assessments to yield a compre-

Metric	Clinically Insignificant Errors		Clinically Significant Errors	
	Kendall's Tau↑ (P-Value↓)	Spearman↑ (P-Value↓)	Kendall's Tau↑ (P-Value↓)	Spearman↑ (P-Value↓)
BLEU-1 (Papineni et al. 2002)	0.348 (8.90e-12)	0.482 (4.87e-13)	0.362 (3.15e-13)	0.513 (8.16e-15)
ROUGE-L (Lin 2004)	0.410 (1.19e-15)	0.558 (8.63e-18)	0.452 (1.14e-19)	0.624 (5.19e-23)
METEOR (Banerjee and Lavie 2005)	0.362 (2.03e-12)	0.500 (4.61e-14)	0.495 (5.00e-23)	0.667 (4.05e-27)
BertScore (Zhang et al. 2019)	0.236 (3.66e-6)	0.339 (8.95e-07)	0.265 (1.03e-7)	0.400 (4.32e-09)
BLUERT (Jiang et al. 2022)	0.329 (9.55e-11)	0.461 (6.72e-12)	0.348 (2.26e-12)	0.493 (1.20e-13)
MOVERScore (Zhao et al. 2019)	-0.353 (4.41e-12)	-0.490 (1.78e-13)	-0.439 (9.29e-19)	-0.603 (3.19e-21)
RadGraphF1 (Jain et al. 2021)	0.374 (7.74e-13)	0.500 (4.49e-14)	0.551 (2.17e-27)	0.717 (6.47e-33)
RaTEScore (Zhao et al. 2024)	0.419 (2.87e-16)	0.563(3.80e-18)	0.507(3.23e-24)	0.682(1.17e-28)
Green (Ostmeier et al. 2024)	0.450 (1.07e-16)	0.596 (1.31e-20)	0.647 (1.68e-34)	0.811 (6.61e-48)
GEMA-Score (Claude-opus-4)	0.379 (1.68e-12)	0.498 (6.11e-14)	0.651 (2.25e-38)	0.822 (2.53e-50)
GEMA-Score (Claude-sonnet-4)	0.361 (1.42e-11)	0.475 (1.89e-12)	0.621 (3.43e-35)	0.799 (1.56e-45)
GEMA-Score (Gemini-2.5-pro)	0.351 (5.89e-11)	0.458 (8.31e-12)	0.643 (1.31e-37)	0.817 (3.87e-49)
GEMA-Score (Deepseek-v3)	0.381 (1.36e-12)	0.500 (4.93e-14)	0.661 (1.45e-39)	0.832 (1.81e-52)
GEMA-Score (Deepseek-r1)	0.361 (2.39e-11)	0.468 (2.69e-12)	0.639 (2.84e-37)	0.816 (5.66e-49)
GEMA-Score (Chat-GPT-4o)	0.367 (8.07e-12)	0.480 (6.21e-13)	0.632 (2.47e-36)	0.810 (9.98e-48)
GEMA-Score (Chat-GPT-o1)	0.372 (3.49e-12)	0.477 (9.26e-13)	0.627 (8.67e-36)	0.803 (2.28e-46)
GEMA-Score (Chat-GPT-o3)	0.371 (4.48e-12)	0.389 (1.95e-13)	0.672 (8.24e-41)	0.846 (6.58e-56)
GEMA-Score (Distilled LLaMA-3.1-8B)	**0.465 (2.36e-17)**	**0.586 (8.14e-20)**	**0.678 (9.16e-41)**	**0.845 (8.48e-56)**

Table 1: Clinical Significance: Human Correlation Comparison of Evaluation Metrics on ReXVal Dataset

Metric	Clinically Insignificant Errors		Clinically Significant Errors	
	Kendall's Tau↑ (P-Value↓)	Spearman↑ (P-Value↓)	Kendall's Tau↑ (P-Value↓)	Spearman↑ (P-Value↓)
BLEU-1 (Papineni et al. 2002)	0.122 (0.111)	0.160 (0.112)	0.147 (0.048)	0.195 (0.052)
ROUGE-L (Lin 2004)	0.104 (0.175)	0.130 (0.197)	0.201 (0.007)	0.259 (0.009)
METEOR (Banerjee and Lavie 2005)	0.114 (0.134)	0.148 (0.143)	0.154 (0.039)	0.208 (0.038)
BertScore (Zhang et al. 2019)	-0.024 (0.754)	-0.029 (0.777)	-0.101 (0.175)	-0.127 (0.207)
BLUERT (Jiang et al. 2022)	0.100 (0.191)	0.134 (0.183)	0.090 (0.225)	0.126 (0.212)
MOVERScore (Zhao et al. 2019)	-0.084 (0.272)	-0.111 (0.271)	-0.005 (0.946)	-0.005 (0.963)
RadGraphF1 (Jain et al. 2021)	0.135 (0.078)	0.179 (0.075)	0.130 (0.082)	0.183 (0.069)
RaTEScore (Zhao et al. 2024)	0.165 (0.031)	0.214 (0.032)	0.177 (0.018)	0.238(0.017)
Green (Ostmeier et al. 2024)	0.118 (0.134)	0.141 (0.163)	0.347 (6.46e-6)	0.433 (6.99e-6)
GEMA-Score (Claude-opus-4)	0.143 (0.075)	0.175 (0.081)	0.423 (1.38e-08)	0.551(2.86e-09)
GEMA-Score (Claude-sonnet-4)	0.173 (0.032)	0.216 (0.031)	0.416 (2.50e-08)	0.542(5.71e-09)
GEMA-Score (Gemini-2.5-pro)	0.154 (0.056)	0.190 (0.059)	0.399 (8.94e-08)	0.521(2.71e-08)
GEMA-Score (Deepseek-v3)	0.173 (0.032)	0.209 (0.037)	0.465 (4.39e-10)	0.608(1.95e-11)
GEMA-Score (Deepseek-r1)	0.185 (0.022)	0.226 (0.024)	0.424 (1.28e-08)	0.549(3.21e-09)
GEMA-Score (Chat-GPT-4o)	0.158 (0.050)	0.190 (0.057)	0.452 (1.31e-09)	0.573(4.59e-10)
GEMA-Score (Chat-GPT-o1)	0.185 (0.021)	0.229 (0.021)	0.466 (3.98e-10)	0.597(5.50e-11)
GEMA-Score (Chat-GPT-o3)	0.176 (0.029)	0.219 (0.029)	0.444 (2.51e-09)	0.573(4.77e-10)
GEMA-Score (Distilled LLaMA-3.1-8B)	**0.133 (0.102)**	**0.167 (0.097)**	**0.414 (3.07e-8)**	**0.533 (1.15e-8)**

Table 2: Clinical Significance: Human Correlation Comparison of Evaluation Metrics on RadEvalX Dataset

hensive evaluation of the generated report:

$$\text{GEMA-Score}(x, \hat{x}) = \alpha \cdot \text{S}_{\text{obj}} + (1 - \alpha) \cdot \text{S}_{\text{sub}}, \quad (6)$$

where $\alpha \in [0, 1]$ balances clinical accuracy and linguistic expressiveness. We set $\alpha = 0.8$ by default to emphasize factual correctness.

Unlike prior metrics that output only scalar scores, our score evaluation agent additionally provides structured outputs following a predefined format, including entity-level false predictions, omissions, and detailed textual explanations for each clinical aspect (e.g., location, severity, uncertainty). This enhances the transparency and interpretability of the overall evaluation process.

Experiment and Analysis

Dataset and Evaluation Settings In this study, we conduct experiments on five datasets: MIMIC-CXR (Johnson et al. 2019), ReXVal (Yu et al. 2022, 2023; Goldberger et al. 2000), RadEvalX (Calamida et al. 2024, 2023; Goldberger et al. 2000), and CT-RATE (Hamamci et al. 2024). MIMIC-CXR includes over 377,000 chest X-rays and 227,000 reports, from which 3,000 cases are selected for correlation analysis. ReXVal contains 50 studies annotated by six radiologists, identifying clinically significant and insignificant errors across six categories. RadEvalX includes 100 studies with similar annotations from two radiologists, covering eight error types. Both datasets are used to assess alignment between automatic scores and expert evaluations. To evaluate the generalizability of GEMA-Score beyond X-rays, we further test it on chest CT data using the CT-RATE dataset. We randomly selected 60 studies with paired reference and generated reports, and conducted human expert annotation of clinical errors with a focus on entity names and locations.

The LLM-based agents operate under deterministic decoding settings, with temperature set to 0 and top-p set to 1, ensuring consistent and reproducible outputs. A maximum token limit of 8192 is used to handle long-form clinical inputs and structured outputs.

Experimental Results

Case Study Analysis Fig. 3 presents three representative case studies to demonstrate how GEMA-Score offers clinically aligned evaluations, especially in both X-ray and

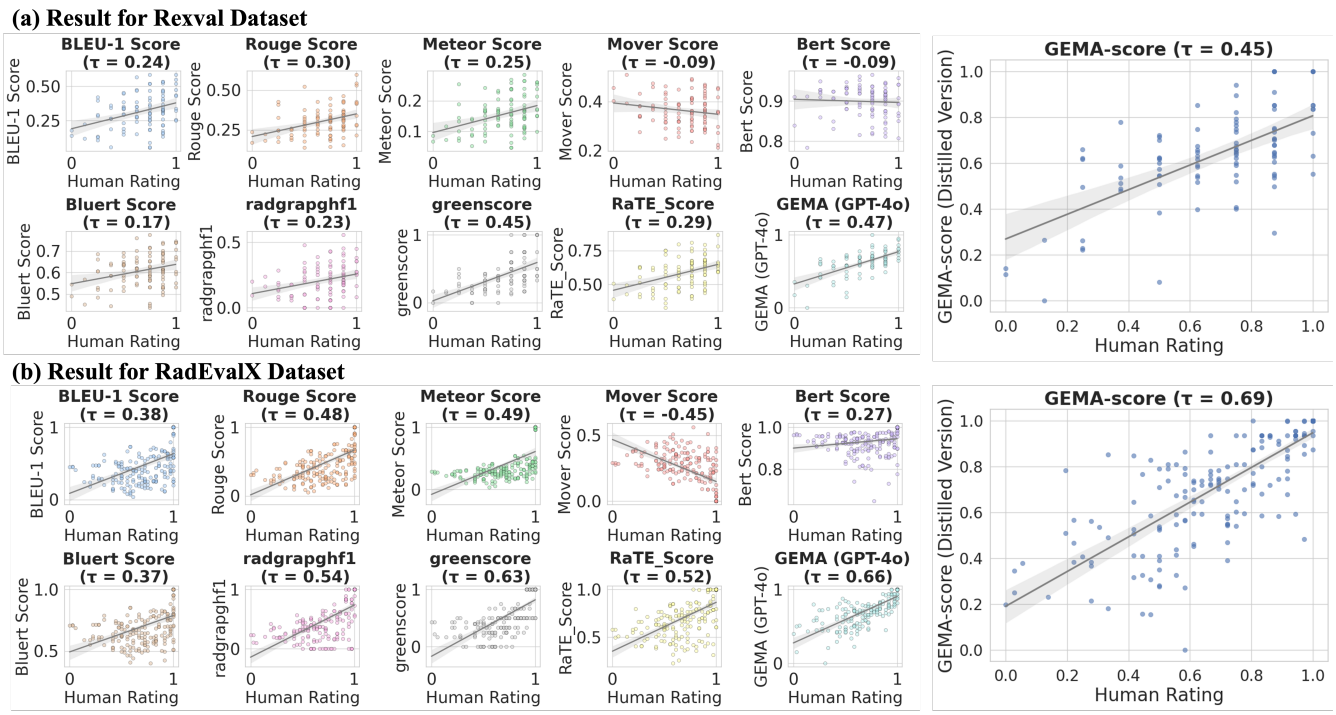

Figure 4: Correlation coefficients with radiologists. (a) Comparison against other metrics on the ReXVal dataset. (b) Comparison against other metrics on the RadEvalX dataset. The rightmost panel is the result of distilled GEMA-score(LLaMA-3.1-8B).

	Overlap-Based			BERT-Based			NER-F1	Chexbert-Based			LLM-based
Method	BLEU-1	ROUGE	METEOR	BERTScore	BLUERT	MOVERScore	RadGraph-F1	P	R	F_1	Green
kendall	0.189	0.177	0.200	0.137	0.211	-0.183	0.054	0.025	0.009	0.021	0.326
p-value	0.006	0.011	0.004	0.050	0.002	0.008	0.436	8.45e-01	9.42e-01	8.70e-01	0.001

Table 3: Correlation and significance between GEMA-Score predictions and different medical report evaluation metrics.

CT modalities. In Case #1914 (X-ray), the generated report omits "diffuse bilateral interstitial opacities", a clinically significant finding; yet traditional metrics (BLEU, METEOR, RadGraph-F1) assign moderate scores (0.25–0.5). In contrast, GEMA-Score assigns a lower score (0.31), consistent with expert annotations highlighting the omission as a major diagnostic error. In Case #1002 (CT), the generated report misses multiple findings related to vertebral changes and degenerative disease. Despite partial semantic matches, diagnostic-based methods yield misleadingly lenient errors = 7. GEMA-Score instead penalizes the omissions and misclassifications with a lower score (0.23), better reflecting the expert-rated 15 significant errors. Overall, GEMA-Score demonstrates stronger alignment with radiologist assessments. Detailed scores and additional examples are provided in the supplementary materials.

Benchmarking Metrics Against Human Expert Judgments Tables 1 and 2 compare the human correlation of different evaluation metrics in identifying clinically significant errors. GEMA-Score consistently outperforms existing metrics across both datasets. While the LLM-based Green score achieves strong correlation with expert ratings (e.g., Spearman = 0.816), GEMA-Score achieves even higher agreement (Spearman = 0.845). Notably, we benchmarked GEMA-Score across a wide range of LLM backbones and APIs, including Claude, Gemini, Chat-GPT, and Deepseek (Kendall's τ: 0.622–0.678 on ReXVal). In addition, we introduce a distilled version of GEMA-Score based on LLaMA-3.1-8B, which achieves satisfactory correlation with human experts on both datasets (e.g., Spearman: 0.845 / 0.533). This demonstrates that GEMA-Score remains robust and effective across model choices while supporting open-source deployment. Fig 4 our score performs the best in correlation analysis across both datasets when combining both error types. Table 3 shows the correlation between GEMA-Score and existing metrics. Among them, Green shows the highest correlation ($\tau = 0.326$), yet a significant difference remains ($p = 0.001$), suggesting that GEMA-Score captures distinct aspects aligned with expert evaluations.

Analysis and Discussion

Consistency Analysis for Expert-Level Evaluation Table 4 reports Pearson correlations between GEMA-Score predictions and ReXVal expert annotations on clinically significant error counts. GEMA† (Chat-GPT-4o) achieves an

	0	1	2	3	4	5	GEMA†	GEMA‡
0	–	0.554	0.737	0.703	0.742	0.617	0.609	0.636
1	0.554	–	0.591	0.603	0.681	0.517	0.457	0.495
2	0.737	0.591	–	0.745	0.730	0.793	0.621	0.677
3	0.703	0.603	0.745	–	0.747	0.644	0.640	0.670
4	0.742	0.681	0.730	0.747	–	0.612	0.655	0.725
5	0.617	0.517	0.793	0.644	0.612	–	0.530	0.582
GEMA†	0.609	0.457	0.621	0.640	0.655	0.530	–	0.776
GEMA‡	0.636	0.495	0.677	0.670	0.725	0.582	0.776	–

† GEMA-Score (Chat-GPT-4o) ‡ GEMA-Score (Distilled LLaMA-3.1)

Table 4: Pearson correlation between GEMA-Score predictions and ReXVal expert annotations across different expert raters (0–5). Higher values indicate stronger agreement.

Rater	(a)	(b)	(c)	(d)
Expert 0	0.468 (4.00e-23)	0.470 (2.47e-23)	0.461 (1.76e-22)	0.462 (1.39e-22)
Expert 1	0.526 (7.99e-30)	0.520 (4.41e-29)	0.424 (7.13e-19)	0.387 (1.06e-15)
Expert 2	0.600 (2.03e-40)	0.582 (1.36e-37)	0.539 (1.52e-31)	0.561 (1.49e-34)
Expert 3	0.645 (1.70e-48)	0.590 (6.84e-39)	0.550 (5.24e-33)	0.514 (2.30e-28)
Expert 4	0.529 (3.17e-30)	0.603 (6.90e-41)	0.506 (2.30e-27)	0.415 (4.68e-18)
Expert 5	0.563 (9.16e-35)	0.497 (2.68e-26)	0.395 (2.11e-16)	0.426 (4.75e-19)
GEMA†	0.591 (5.58e-19)	0.487 (1.58e-12)	0.507 (1.83e-14)	0.563 (4.08e-18)
GEMA‡	0.651 (2.29e-25)	0.512 (1.15e-14)	0.520 (2.97e-15)	0.503 (3.22e-14)

† GEMA-Score (Chat-GPT-o1) ‡ GEMA-Score (Distilled LLaMA-3.1)

Table 5: Pearson correlation between GEMA-Score predictions and ReXVal expert to the mean rater. Error types: (a) false positive finding, (b) missed reference finding, (c) misidentified location, (d) incorrect severity.

average correlation of 0.585, approaching the average inter-expert agreement (0.668). The distilled version, GEMA‡ (LLaMA-3.1), yields an even higher consistency with a mean correlation of 0.630, demonstrating its reliability in a lightweight setting. Table 5 further analyzes model behavior across four error types. The distilled GEMA‡ achieves the highest average correlation with experts across all categories, reaching 0.651 for false positives (a), 0.512 for missed findings (b), 0.520 for location errors (c), and 0.503 for severity errors (d). These results indicate strong alignment with expert fine-grained judgments. Further, the distilled GEMA‡ yields consistently higher agreement, validating its reliability for fine-grained clinical assessment.

Distilled Model Evaluation for Local Deployment Table 6 shows that LLaMA-3.1-8B achieves the best performance, with the highest correlation on ReXVal for both significant (0.678) and insignificant errors (0.465), and strong performance on RadEvalX (0.414 / 0.133). Phi-2-2.7B offers comparable results (0.676 / 0.459 on ReXVal and 0.415 / 0.131 on RadEvalX), making it an alternative for resource-constrained environments. Qwen-2VL-2B lags behind with lower correlation values, especially for ReXVal significant errors (0.648) and RadEvalX significant errors (0.369).

Generalization Analysis for CT Report Evaluation Table 7 presents GEMA-Score performance on CT reports

Distilled Model	**ReXVal**		**RadEvalX**	
	Sig. Corr. ↑	Insig. Corr. ↑	Sig. Corr. ↑	Insig. Corr. ↑
LLaMA-3.1-8B (Touvron et al. 2023)	0.678 (9.16e-41)	0.465 (2.36e-17)	0.414 (3.07e-8)	0.133 (0.102)
Phi-2-2.7B (Javaheripi et al. 2023)	0.676 (1.90e-40)	0.459 (1.89e-16)	0.415 (6.67e-9)	0.131 (0.024)
Qwen-2VL-2B (Team 2024)	0.648 (4.86e-37)	0.422 (4.60e-14)	0.369 (8.94e-7)	0.221 (0.008)

Table 6: Correlation on significant and insignificant findings across ReXVal and RadEvalX with different distilled models.

Agent Setting	**F1 Score**	
	Kendall (p) ↑	Spearman (p) ↑
BLEU-1 (Papineni et al. 2002)	0.231 (2.3e-02)	0.291 (2.4e-02)
ROUGE-L (Lin 2004)	0.218 (3.1e-02)	0.281 (3.0e-02)
BLUERT (Jiang et al. 2022)	0.226 (2.5e-02)	0.285 (2.7e-02)
RadGraphF1 (Jain et al. 2021)	0.224 (2.7e-02)	0.288 (2.6e-02)
Radbert (Yan et al. 2022)	0.170 (1.0e-01)	0.209 (1.1e-01)
GREEN (Ostmeier et al. 2024)	0.310 (2.6e-03)	0.381 (2.7e-03)
GEMA-Score (Deepseek-V3)	0.576 (1.7e-07)	0.632 (7.9e-08)
GEMA-Score (Deepseek-R1)	0.482 (3.8e-06)	0.526 (1.8e-05)
GEMA-Score (Chat-GPT-4o)	0.592 (2.8e-08)	0.668 (5.2e-09)
GEMA-Score (Chat-GPT-o1)	0.755 (9.7e-13)	0.794 (3.8e-14)
GEMA-Score (Chat-GPT-o3)	0.712 (1.6e-11)	0.774 (4.0e-13)
GEMA-Score (LLaMA-3.1-8B)	0.660 (1.0e-09)	0.738 (1.7e-11)
GEMA-Score (Phi-2-2.7B)	0.341 (1.9e-03)	0.377 (2.9e-03)
GEMA-Score (Qwen-2VL-2B)	0.628 (4.4e-09)	0.708 (2.4e-10)
GEMA-Score (LLaMA-3.1&CoT)	0.689 (5.4e-11)	0.741 (1.2e-11)

Table 7: Assessment of F1 Score Consistency on CT Reports.

from the CT-RATE dataset, demonstrating its generalizability. GEMA-Score consistently outperforms the single-step baseline GREEN across all metrics (e.g., Kendall correlation from 0.310 to 0.755). The best results are achieved by Chat-GPT-o1 (Struct. Error: 0.733±1.191), with strong performance also observed from Chat-GPT-o3 and Chat-GPT-4o. Among distilled models, LLaMA-3.1-8B performs best (Kendall: 0.660, Spearman: 0.738), and the LLaMA-3.1+Chain-of-Thought variant achieves further improvements (Kendall: 0.689, Spearman: 0.741) while yielding the lowest structural error (0.633±1.301). These results underscore the value of multi-agent, multi-step reasoning for robust clinical report evaluation across imaging modalities.

Conclusion

GEMA-Score offers a structured multi-agent framework that assesses both objective clinical accuracy and subjective report quality. It aligns well with expert judgments, validating its clinical reliability. Further, we introduce a distilled version for local deployment and extend its use from chest X-rays to CT reports, demonstrating cross-modality generalizability. These advances strengthen GEMA-Score as a scalable and trustworthy tool for medical report evaluation.

Acknowledgements

Guang Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, NIHR Imperial Biomedical Research Centre (RDA01), The Wellcome Leap Dynamic resilience program (co-funded by Temasek Trust)., UKRI guarantee funding for Horizon Europe MSCA Postdoctoral Fellowships (EP/Z002206/1), UKRI MRC Research Grant, TFS Research Grants (MR/U506710/1), Swiss National Science Foundation (Grant No. 220785), and the UKRI Future Leaders Fellowship (MR/V023799/1, UKRI2738). Zhenxuan Zhang was supported by a CSC Scholarship.

References

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Boag, W.; Kané, H.; Rawat, S.; Wei, J.; and Goehler, A. 2021. A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 458–465. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Calamida, A.; Nooralahzadeh, F.; Rohanian, M.; Fujimoto, K.; Nishio, M.; and Krauthammer, M. 2023. Radiology-Aware Model-Based Evaluation Metric for Report Generation. arXiv:2311.16764.

Calamida, A. R.; Nooralahzadeh, F.; Rohanian, M.; Nishio, M.; Fujimoto, K.; and Krauthammer, M. 2024. Radiology Report Generation Models Evaluation Dataset For Chest X-rays (RadEvalX) (version 1.0.0).

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Estornell, A.; and Liu, Y. 2024. Multi-LLM Debate: Framework, Principals, and Interventions. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 28938–28964. Curran Associates, Inc.

Goldberger, A. L.; Amaral, L. A. N.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220. [Online].

Gu, B.; Desai, R. J.; Lin, K. J.; and Yang, J. 2024. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1): 367.

Hamamci, I. E.; Er, S.; Wang, C.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Doga, I.; Durugol, O. F.; Dai, W.; Xu, M.; et al. 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C. C. T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3): 3.

Jiang, H.; Zhang, C.; Xin, Z.; Huang, X.; Li, C.; and Tai, Y. 2022. Transfer learning based on lexical constraint mechanism in low-resource machine translation. *Computers and Electrical Engineering*, 100: 107856.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6: 317.

Li, X.; Wang, S.; Zeng, S.; Wu, Y.; and Yang, Y. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1): 9.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Michalson, A. E.; Moseley, M.; Langlotz, C.; Chaudhari, A. S.; et al. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. *arXiv preprint arXiv:2405.03595*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2023. Medhalt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv:2004.09167.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, S.; Long, Z.; Fan, Z.; Huang, X.; and Wei, Z. 2025. Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3310–3328. Abu Dhabi, UAE: Association for Computational Linguistics.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Xie, Y.; Zhang, S.; Cheng, H.; Liu, P.; Gero, Z.; Wong, C.; Naumann, T.; Poon, H.; and Rose, C. 2024. DocLens: Multi-aspect Fine-grained Evaluation for Medical Text Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 649–679. Bangkok, Thailand: Association for Computational Linguistics.

Yan, A.; McAuley, J.; Lu, X.; Du, J.; Chang, E. Y.; Gentili, A.; and Hsu, C.-N. 2022. RadBERT: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4): e210258.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Ho Lee, H. M.; Abad, Z. S. H.; Ng, A. Y.; Langlotz, C. P.; Venugopal, V. K.; and Rajpurkar, P. 2022. Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation. *medRxiv*.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Kaiser Ururahy Nunes Fonseca, E.; Lee, H.; Shakeri, Z.; Ng, A.; Langlotz, C.; Venugopal, V. K.; and Rajpurkar, P. 2023. Radiology Report Expert Evaluation (ReXVal) Dataset (version 1.0.0).

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A Metric for Radiology Report Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15004–15019. Miami, Florida, USA: Association for Computational Linguistics.