RADAR: A Multimodal Benchmark for 3D Image-Based Radiology Report Review

Zhaoyi Sun[1], Minal Jagtiani[1], Wen-wai Yim[2], Fei Xia[1], Martin Gunn[1], Meliha Yetisgen[1†], and Asma Ben Abacha[2†]

[1] University of Washington, Seattle, WA 98195, USA
{zhaoyis,minalj,fxia,marting,melihay}@uw.edu
[2] Microsoft Health AI, Redmond, WA 98052, USA
{yimwenwai,abenabacha}@microsoft.com

Abstract. Radiology reports for the same patient examination may contain clinically meaningful discrepancies arising from interpretation differences, reporting variability, or evolving assessments. Systematic analysis of such discrepancies is important for quality assurance, clinical decision support, and multimodal model development, yet remains limited by the lack of standardized benchmarks. We present RADAR, a multimodal benchmark for radiology report discrepancy analysis that pairs 3D medical images with a preliminary report and corresponding candidate edits for the same study. The dataset reflects a standard clinical workflow in which trainee radiologists author preliminary reports that are subsequently reviewed and revised by attending radiologists. RADAR defines a structured discrepancy assessment task requiring models to evaluate proposed edits by determining image-level agreement, assessing clinical severity, and classifying edit type (correction, addition, or clarification). In contrast to prior work emphasizing binary error detection or comparison against fully independent reference reports, RADAR targets fine-grained clinical reasoning and image–text alignment at the report review stage. The benchmark consists of expert-annotated abdominal CT examinations and is accompanied by standardized evaluation protocols to support systematic comparison of multimodal models. RADAR provides a clinically grounded testbed for evaluating multimodal systems as reviewers of radiology report edits.

Keywords: Radiology report evaluation · Multimodal benchmark · VLM.

1 Introduction

Radiology report discrepancies are a recognized concern for patient safety and quality assurance, occurring in approximately 3–5% of routine interpretations, with higher rates reported in emergency department (ED) imaging [4]. In ED settings, particularly within academic centers where preliminary resident reports are later reviewed by attendings, even small discrepancy rates can have disproportionate downstream impact, as clinical decisions may be made before the

† These authors contributed equally as senior authors.

final interpretation is available [6, 19]. Large retrospective studies have reported approximately 2.6% of emergency imaging interpretations contained discrepancies requiring direct physician notification due to potential impact on patient care, with contrast-enhanced abdominal CT among the most commonly involved examinations [19]. Although many discrepancies do not alter management, a clinically meaningful subset is associated with negative or significant effects on patient care, underscoring the need for systematic discrepancy analysis.

Despite rapid progress in medical AI, existing datasets and benchmarks for radiology discrepancy or error analysis remain limited, particularly for volumetric CT. Many prior approaches consider errors as text-only phenomena or rely on synthetically introduced errors (e.g., word insertion, deletion, or substitution), which may not faithfully represent clinically meaningful, image-evidenced interpretive discrepancies and are often not detectable without true image understanding [22, 25]. Meanwhile, large-scale 3D CT vision–language datasets and benchmarks developed for report generation or visual question answering (VQA) are not designed around the real-world preliminary-to-final reporting workflow and do not directly evaluate whether a proposed edit is supported by imaging evidence or resolves a discrepancy. To the best of our knowledge, this work presents the first multimodal benchmark[1] specifically designed for image-grounded radiology report discrepancy analysis that conditions on a CT image, a preliminary report, and a candidate suggested edit. Our contributions are as follows:

- A multimodal benchmark, RADAR, derived from real trainee-to-attending report revisions, rather than synthetically introduced errors, enabling evaluation of whether candidate edits are supported by imaging evidence. The RADAR dataset was recently used in the MEDIQA-Core-Task-2 2026 challenge[2] [2].
- A fine-grained evaluation framework that jointly assesses agreement, clinical severity, and discrepancy type.
- An empirical study of several vision–language foundation models under multiple volumetric input settings, establishing a baseline for future methods on image-grounded discrepancy analysis.

2 Related Work

2.1 Quality Assurance and Error Detection

Radiology report discrepancies have been widely studied in the context of quality assurance, education, and clinical risk mitigation. Lyo et al. [15] analyzed preliminary-to-final report revisions and showed that generative models can convert real-world edits into structured educational feedback, highlighting the value

[1] The benchmark will be made available upon request and completion of a Data Use Agreement (DUA), in compliance with institutional and hospital data governance policies. More information about requesting access to the dataset and completing the DUA can be found at the following link: https://github.com/GeraldSun/RADAR

[2] https://www.imageclef.org/2026/medical/mediqa-core

of discrepancy analysis. Recent studies have extended this perspective to multimodal error detection by incorporating imaging data. Wu et al. [24] evaluated multimodal LLMs on a benchmark constructed by injecting synthetic insertion, removal, and substitution errors into radiology reports paired with images; although fine-tuned models achieved clinician-level performance in binary detection, they struggled with fine-grained classification. MedErr-CT introduced a CT-specific VQA benchmark and evaluated several 3D multimodal models, reporting substantial performance gaps across error types and task levels [12]. However, these studies rely primarily on synthetic errors and constrained tasks, and do not explicitly evaluate whether report edits are image-supported.

A substantial body of work focuses on text-only error detection. Early NLP approaches modeled insertion, substitution, and deletion errors using sequence-to-sequence or classification architectures, showing that neural models can detect implausible report content [25]. More recently, GPT-based and fine-tuned LLMs have been applied to identify factual inconsistencies and improve reporting accuracy [7, 20, 21]. Other efforts such as GREEN, which introduces structured error annotation and factuality-aware metrics [17], and ReXErr, which synthesizes clinically meaningful errors [18], further standardize report evaluation. Although these text-centric approaches do not verify image evidence, they provide insights into discrepancy characterization and evaluation methodology.

2.2 Vision-Language Models (VLMs) in 3D Radiology

Vision-language modeling in 3D radiology can be grouped by input representation. Volumetric models (e.g., RadFM [23], CT-CLIP [8], Merlin [3], M3D-LaMed [1]) employ native 3D encoders trained with multimodal objectives [23, 8, 3, 1]. Slice-based approaches instead adapt 2D VLM backbones using slice pooling or selection strategies.

Recent medical multimodal and generalist models (e.g., Lingshu [13], Hulu-Med [10], Gemini-3-pro, GPT-5.2, Qwen3.5-plus) operate on sampled slices rather than volumetric tokens [13, 10]. However, cross-paradigm comparisons remain limited. Many studies do not benchmark against strong slice-based baselines under matched settings, and reported gains often reflect data scale and supervision rather than architecture alone. The relative advantages of native 3D encoding versus slice-based prompting therefore remain unclear.

2.3 Datasets for 3D Radiology Report Generation and VQA

Progress in 3D CT vision–language modeling has been driven by the release of datasets pairing volumetric scans with reports. CT-RATE, comprising 25,692 non-contrast 3D chest CT scans with reports, has become a foundational benchmark for models such as CT-CLIP and CT-CHAT [8]. RadGenome-Chest CT augments CT-RATE with organ-level segmentation masks, sentence-level grounding annotations, and grounded VQA pairs, enabling regionally grounded reasoning and fine-grained evaluation [26]. Despite these advances, publicly available

3D CT datasets remain limited in scale and diversity compared to 2D radiography benchmarks, constraining systematic evaluation of volumetric VLMs.

Alongside dataset development, several studies have explored 3D CT report generation and VQA. Li et al. [14] proposed a brain CT report generation framework and introduced FORTE, a feature-oriented evaluation scheme designed to better reflect diagnostic quality than text-overlap metrics. Other approaches incorporate explicit anatomical or spatial modeling, such as CT-GRAPH, which encodes anatomical hierarchies for report generation on CT-RATE [11], and 3D-CT-GPT, which frames report generation through VQA-style conditioning [5]. In 3D CT VQA, CT-Agent adopts an agentic framework to address long-range spatial dependencies across slices [16], while M3 extends multimodal modeling to abdominal CT [9]. Although these efforts focus on generation and question answering rather than discrepancy resolution, they highlight challenges in volumetric reasoning that motivate more targeted benchmarks.

3 Task Definition

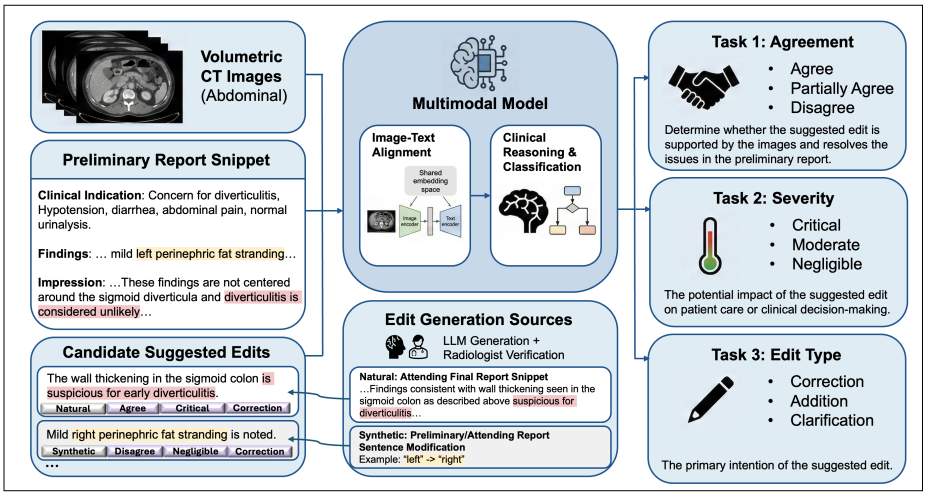

Fig. 1: Overview of the RADAR workflow.

Given volumetric CT images, a preliminary report, and a candidate suggested edit referring to the same study, the task is to evaluate whether the suggested edit is supported by the imaging evidence and resolves a discrepancy in the preliminary report. Specifically, models are required to: (1) determine the level of image-grounded agreement of the suggested edit (agree, partially agree, or disagree), (2) assess the clinical severity of the underlying discrepancy (critical, moderate, or negligible), and (3) classify the edit type as a correction, addition,

Table 1: Dataset characteristics and annotation statistics. *Natural* denotes discrepancies arising from preliminary-to-attending report revisions. *Mixed* includes natural discrepancies and disagree edits introduced for evaluation balancing.

(a) Case-level Statistics

Number of Cases	
Total CT examinations	50
Validation cases	20
Test cases	30
Daytime cases	25
Overnight cases	25
Report Length (words), mean (SD)	
Preliminary	
Daytime	220.48 (66.74)
Overnight	88.00 (38.28)
Imaging Statistics	
Image series	319
Slices	45,838

(b) Edit-level Statistics

	Natural		**Mixed**	
Count	**Dev**	**Test**	**Dev**	**Test**
Total candidate edits	102	153	184	274
Agreement				
Agree	88	132	88	132
Partially Agree	11	16	11	16
Disagree	3	5	85	126
Severity				
Negligible	67	117	94	141
Moderate	25	27	45	69
Critical	10	9	45	64
Edit Type				
Addition	61	110	108	192
Correction	27	28	42	51
Clarification	14	15	34	31

or clarification. The task is formulated to reflect real-time report review, where models evaluate candidate edits without access to the finalized attending report. Figure 1 illustrates the RADAR workflow and task structure.

4 Dataset Creation

4.1 Case Selection and Data Sources

We collected abdomen and pelvis CT examinations from the ED at Harborview Medical Center (HMC), UW Medicine. Each examination includes all CT image series (DICOM format) along with paired preliminary and attending final radiology reports. Preliminary reports are authored by residents or fellows and are available continuously (24/7), while attending radiologists subsequently issue final reports after reviewing both the imaging and the preliminary interpretation. This study was conducted under Institutional Review Board (IRB) approval.

The dataset includes both daytime and overnight ED cases, where the designation (daytime vs. overnight) refers to the time at which the preliminary report was authored. Daytime preliminary reports are typically more comprehensive and later refined by attendings, whereas overnight reports often follow a "quick-look" workflow in which residents focus on identifying acute findings requiring immediate attention, with attendings subsequently expanding or correcting the interpretation. Table 1(a) shows case-level statistics of RADAR. All data were de-identified. We treat the attending final report as the reference standard for identifying discrepancies relative to the preliminary report.

4.2 Preprocessing and Candidate Edit Generation

Candidate suggested edits were derived by first identifying discrepancies between preliminary and final reports using GPT-4o, followed by manual review by an attending radiologist. The radiologist rejected incorrect suggestions, added missed discrepancies, and reformulated each verified discrepancy into a concise edit intended to resolve the difference in the preliminary report.

Edits referring to longitudinal trends or prior examinations were excluded, as they cannot be verified from a single CT study.

4.3 Gold-Standard Annotation and Dataset Balancing

Each candidate edit was annotated by a radiologist with 10 years of post-board-certification experience along three dimensions: *agreement*, *severity*, and *edit type*. *Agreement* indicates whether the edit is factually accurate and supported by the imaging evidence (agree), generally correct but minor or redundant (partially agree), or incorrect/unnecessary (disagree). *Severity* reflects the potential clinical impact relative to the indication (critical, moderate, or negligible). *Edit type* captures the primary intention of the edit (correction, addition, or clarification) regardless of its accuracy. Dataset statistics are summarized in Table 1.

We observed substantial class imbalance in the natural dataset (220 agree, 27 partially agree, 8 disagree). To support balanced evaluation, we introduced a set of synthetic disagree edits used solely to balance the evaluation split and assess model robustness to unsupported edits. These were generated using GPT-5.2 to produce plausible but incorrect modifications of otherwise valid report statements, guided by previously described error patterns (e.g., ReXErr [18]) and the empirical edit type distribution of natural discrepancies. All synthetic edits were manually reviewed and annotated by a board-certified radiologist to confirm the correct label as disagree. As summarized in Table 1(b), the *Natural* evaluation set consists exclusively of discrepancies derived from real preliminary-to-attending report revisions, whereas the *Mixed* evaluation set combines these natural discrepancies with the synthetic disagree edits. We report results on both evaluation settings to assess performance under realistic clinical conditions as well as under a more balanced agreement-label distribution.

5 Methods

5.1 Baseline Systems

We evaluate several multimodal foundation models on RADAR. Prompts, preprocessing steps, and inference configurations are included in released codebase.

- **Series Selection.** Because CT studies contain multiple series, we first select the most relevant series for each candidate edit using GPT-OSS-20B. The model receives structured DICOM metadata (e.g., anatomical region, acquisition timing) and the edit text, and outputs the best-matched series (e.g., COR BODY). Subsequent inputs are restricted to the selected series.

- **Gemini-2.5-Pro and Gemini-3-Pro.** Multimodal models developed by Google. Evaluated with four visual input settings: 50, 20, or 10 uniformly sampled slices, and a video constructed from all slices.
- **Qwen3.5-plus.** Multimodal model developed by Alibaba. Evaluated under the same four slice/video configurations for direct comparison.

5.2 Evaluation Metrics

We report metrics under two evaluation settings that differ only in the test subset. Accuracy is reported for agreement, severity, and edit type. We additionally compute two *Composite Scores* to jointly assess agreement and fine-grained reasoning. Unless otherwise stated, we use the *Relaxed Composite Score* as the primary metric for model comparison and analysis. For each instance, agreement is considered correct under a fuzzy matching criterion in which *Agree* and *Partially Agree* are treated as equivalent, while *Disagree* remains distinct. If the predicted and gold agreement labels do not match under this criterion, the instance receives a score of 0. If agreement matches, the instance receives a score of 1 only when both the severity label and the edit-type label exactly match the gold annotations; otherwise, the score is 0. The Relaxed Composite Score is averaged across all instances and reflects end-to-end performance on discrepancy analysis. We also report a *Strict Composite Score*, which follows the same definition except that agreement must match exactly across all three agreement categories. The evaluation code is available on our project's GitHub repository[3].

We report results on (i) the **Mixed** set (natural edits combined with synthetic disagree edits introduced for evaluation balancing), which evaluates both clinical utility and safety since synthetic edits are constructed to be unsupported by the image evidence, and (ii) the **Natural** subset, which contains real radiologist revisions and reflects performance on real-world corrections without synthetic perturbations. We additionally report runtime, measured as the total time required for each experiment setting.

6 Results

Table 2 includes the performance results. On the Mixed dataset, Edit Type is consistently high across models (0.78–0.84), while Agreement (0.46–0.70) and Severity (0.46–0.56) show moderate variation. However, Composite Score remains comparatively low across settings (0.19–0.40), reflecting the difficulty of jointly satisfying agreement, severity, and edit type prediction within a single instance. For Gemini-3-Pro, 20 slices achieves the highest Agreement (0.693), while 50 slices yields the highest Composite Score (0.391). Similarly, Qwen3.5-plus does not exhibit monotonic gains from increasing slice counts; its best Composite Score occurs at 20 slices (0.255), with 50 slices slightly declining (0.245). Video-style input does not consistently improve results: it benefits Qwen3.5-plus

[3] https://github.com/GeraldSun/RADAR

Table 2: Baseline performance on RADAR. All values are accuracy for Agreement (Agr), Severity (Sev), Edit Type (Typ), Composite Score (Strict; Comp-S), and Composite Score (Relaxed; Comp-R).

Model / Setting	Mixed					Natural					Time (h)
	Agr	Sev	Typ	Comp-S	Comp-R	Agr*	Sev	Typ	Comp-S	Comp-R	
Gemini-2.5-Pro											
10 slices	0.522	0.467	0.814	0.193	0.219	0.791	0.412	0.856	0.301	0.346	1.1
20 slices	0.489	0.464	0.807	0.168	0.190	0.784	0.425	0.837	0.288	0.327	1.2
50 slices	0.460	0.526	0.814	0.248	0.277	0.680	0.562	0.850	0.386	0.438	1.3
video (all)	0.533	0.464	0.814	0.201	0.234	0.797	0.438	0.837	0.314	0.373	1.5
Gemini-3-Pro											
10 slices	0.686	0.536	0.803	0.321	0.361	0.647	0.614	0.830	0.386	0.458	1.8
20 slices	**0.693**	0.544	0.781	0.321	0.361	0.660	0.614	0.791	0.379	0.451	2.0
50 slices	0.675	**0.560**	0.821	**0.339**	**0.391**	0.627	0.614	0.843	**0.399**	**0.484**	3.0
video (all)	0.653	0.558	0.818	0.307	0.347	0.575	**0.647**	0.830	0.353	0.425	2.6
Qwen3.5-plus											
10 slices	0.558	0.533	**0.839**	0.223	0.274	0.418	0.601	**0.869**	0.216	0.307	3.3
20 slices	0.591	0.536	0.818	0.255	0.314	0.477	0.621	0.824	0.255	0.359	3.4
50 slices	0.613	0.518	0.818	0.245	0.288	0.510	0.569	0.830	0.242	0.320	3.7
video (all)	0.628	0.544	**0.839**	0.296	0.350	0.503	0.627	0.863	0.320	0.418	3.5

*Agreement labels in the Natural subset are highly imbalanced across classes.

but does not outperform slice-based settings for Gemini models. On the Natural dataset, Agreement must be interpreted cautiously due to strong class imbalance; for example, Gemini-2.5-Pro attains high Agreement likely influenced by the dominant "Agree" class. Gemini-3-Pro (50 slices) achieves the highest Composite Score (0.484). Overall, performance differences appear to be primarily driven by model family, and neither increasing slice count nor adopting video-style input reliably guarantees improved accuracy across settings.

7 Discussion

Our results highlight distinct challenges across the three components of discrepancy reasoning: agreement detection, severity assessment, and edit type classification. Edit type classification is comparatively easier, as it primarily relies on understanding the semantic function of an edit. In contrast, agreement detection requires robust cross-modal alignment to determine whether textual claims are supported by image evidence, particularly in the presence of synthetic edits designed to be unsupported by image evidence. Severity assessment remains the most challenging, as it demands calibrated clinical judgment to evaluate the practical impact of a discrepancy within real-world workflows. Our initial experiments suggest that while current multimodal models can effectively recognize linguistic patterns, reliable image-grounded and clinically nuanced reasoning remains difficult. A discrepancy-aware model could therefore serve as a verification layer, confirming image-supported corrections while flagging unsupported or clinically impactful edits before they influence downstream systems. Such capability may be particularly valuable in ED and resource-limited settings, where attending review may be delayed, and may ultimately strengthen the safety and robustness of AI-assisted clinical pipelines.

8 Conclusion

We present RADAR, a multimodal benchmark for image-grounded radiology report discrepancy analysis that evaluates agreement detection, severity assessment, and edit type classification within a realistic preliminary-to-attending review workflow. Our results show that while large multimodal models can effectively recognize linguistic edit patterns, reliable cross-modal verification and clinically calibrated severity reasoning remain challenging. Grouded in real preliminary-to-attending revisions, RADAR provides a clinically meaningful testbed for discrepancy-aware reasoning. While currently limited by its modest size and focus on a single modality and body site, RADAR establishes a foundation for broader multimodal discrepancy benchmarks. Future work will expand to additional modalities and anatomical regions and incorporate longitudinal cross-exam reasoning to further improve safety and workflow robustness.

Limitations

Several limitations should be considered. First, RADAR is limited in scale, consisting of abdominal CT examinations collected from a single academic medical center. As a result, the dataset may not capture the full diversity of reporting styles, patient populations, imaging protocols, or discrepancy patterns observed across institutions and subspecialties. Second, naturally occurring disagreement cases were uncommon in the underlying clinical workflow, requiring the introduction of expert-verified synthetic disagree edits for balanced evaluation. Although we report results separately on the Natural and Mixed settings, synthetic edits may not fully reflect the distribution of real-world reporting discrepancies. Third, severity assessment involves clinical judgment and may contain inherent subjectivity despite the use of structured annotation guidelines.

Acknowledgment

We would like to thank the ImageCLEF and MEDIQA-CORE 2026 organizers for their collaboration and support in hosting the shared task based on RADAR. We also thank all participating teams for their valuable contributions, innovative approaches, and engagement, which have helped advance research on multimodal radiology report discrepancy analysis.

Disclosure of Interest

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

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3D medical image analysis with multi-modal large language models. arXiv [cs.CV] (31 Mar 2024)
2. Ben Abacha, A., Sun, Z., Yim, W., Xia, F., Yetisgen, M.: Overview of the mediqa-core 2026 task 2 on radiology report discrepancy assessment. In: CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 2026)
3. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., Bluethgen, C., Jensen, M.E.K., Ostmeier, S., Varma, M., Valanarasu, J.M.J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., Weng, W.H., Amaro, E., Ahuja, N., Fries, J., Shah, N.H., Johnston, A., Boutin, R.D., Wentland, A., Langlotz, C.P., Hom, J., Gatidis, S., Chaudhari, A.S.: Merlin: A vision language foundation model for 3D computed tomography. Res. Sq. (28 Jun 2024)
4. Brady, A.P.: Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging **8**(1), 171–182 (Feb 2017)
5. Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., Guo, L., Han, J., Liu, T., Liu, J., Zhang, T.: 3D-CT-GPT: Generating 3D radiology reports through integration of large vision-language models. arXiv [cs.CV] (28 Sep 2024)
6. Friedman, S.M., Merman, E., Chopra, A.: Clinical impact of diagnostic imaging discrepancy by radiology trainees in an urban teaching hospital emergency department. Int. J. Emerg. Med. **6**(1), 24 (16 Jul 2013)
7. Gertz, R.J., Dratsch, T., Bunck, A.C., Lennartz, S., Iuga, A.I., Hellmich, M.G., Persigehl, T., Pennig, L., Gietzen, C.H., Fervers, P., Maintz, D., Hahnfeldt, R., Kottlors, J.: Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. Radiology **311**(1), e232714 (Apr 2024)
8. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., Dai, W., Xu, M., Reynaud, H., Dasdelen, M.F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Kaplan, A., Lu, Z., Polacin, M., Kainz, B., Bluethgen, C., Batmanghelich, K., Ozdemir, M.K., Menze, B.: Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv [cs.CV] (30 Sep 2025)
9. Hosseini, A., Ibrahim, A., Serag, A.: M3: multimodal artificial intelligence for medical report generation and visual question answering from 3D abdominal CT scans. BJR Artif. Intell. (2025)
10. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., Hao, J., Chen, Z., Wu, R., Tang, T., Lv, J., Xu, H., Wang, H., Xiao, J., Feng, B., Zhu, F., Li, K., Xie, W., Sun, J., Wu, J., Liu, Z.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv [cs.CV] (5 Nov 2025)
11. Kalisch, H., Hörst, F., Kleesiek, J., Herrmann, K., Seibold, C.: CT-GRAPH: Hierarchical graph attention network for anatomy-guided CT report generation. arXiv [cs.CV] (7 Aug 2025)
12. Kyung, S., Park, H., Seo, J., Sung, J., Kim, J., Kim, D., Jo, W., Nam, Y., Park, S., Kwon, T., Lee, S.M., Kim, N.: MedErr-CT: A visual question answering benchmark for identifying and correcting errors in CT reports. arXiv [cs.CV] (24 Jun 2025)

13. LASA Team, Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., Sun, Y., Shen, J., Wang, C., Tan, J., Zhao, D., Xu, T., Zhang, H., Rong, Y.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv [cs.CL] (8 Jun 2025)
14. Li, C.Y., Chang, K.J., Yang, C.F., Wu, H.Y., Chen, W., Bansal, H., Chen, L., Yang, Y.P., Chen, Y.C., Chen, S.P., Chen, S.J., Lirng, J.F., Chang, K.W., Chiou, S.H.: Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat. Commun. **16**(1), 2258 (6 Mar 2025)
15. Lyo, S., Mohan, S., Hassankhani, A., Noor, A., Dako, F., Cook, T.: From revisions to insights: Converting radiology report revisions into actionable educational feedback using generative AI models. J Imaging Inform Med **38**(2), 1265–1279 (Apr 2025)
16. Mao, Y., Xu, W., Qin, Y., Gao, Y.: CT-agent: A multimodal-LLM agent for 3D CT radiology question answering. arXiv [cs.CV] (22 May 2025)
17. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J.B.: GREEN: Generative radiology report evaluation and error notation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.21, https://aclanthology.org/2024.findings-emnlp.21/
18. Rao, V.M., Zhang, S., Acosta, J.N., Adithan, S., Rajpurkar, P.: ReXErr: Synthesizing clinically meaningful errors in diagnostic radiology reports. In: Biocomputing 2025. pp. 70–81. WORLD SCIENTIFIC (21 Dec 2024)
19. Ruchman, R.B., Jaeger, J., Wiggins, 3rd, E.F., Seinfeld, S., Thakral, V., Bolla, S., Wallach, S.: Preliminary radiology resident interpretations versus final attending radiologist interpretations and the impact on patient care in a community hospital. AJR Am. J. Roentgenol. **189**(3), 523–526 (Sep 2007)
20. Salam, B., Stüwe, C., Nowak, S., Sprinkart, A.M., Theis, M., Kravchenko, D., Mesropyan, N., Dell, T., Endler, C., Pieper, C.C., Kuetting, D.L., Luetkens, J.A., Isaak, A.: Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models. Eur. Radiol. (20 Feb 2025)
21. Sun, C., Teichman, K., Zhou, Y., Critelli, B., Nauheim, D., Keir, G., Wang, X., Zhong, J., Flanders, A.E., Shih, G., Peng, Y.: Generative large language models trained for detecting errors in radiology reports. Radiology **315**(2), e242575 (May 2025)
22. Sun, Z., Yim, W.W., Uzuner, Ö., Xia, F., Yetisgen, M.: A scoping review of natural language processing in addressing medically inaccurate information: Errors, misinformation, and hallucination. J. Biomed. Inform. **169**(104866), 104866 (Sep 2025)
23. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nat. Commun. **16**(1), 7866 (23 Aug 2025)
24. Wu, J., Kim, Y., Keller, E.C., Chow, J., Levine, A.P., Pontikos, N., Ibrahim, Z., Taylor, P., Williams, M.C., Wu, H.: Exploring multimodal large language models for radiology report error-checking. arXiv [cs.CL] (20 Dec 2023)
25. Zech, J., Forde, J., Titano, J.J., Kaji, D., Costa, A., Oermann, E.K.: Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Ann. Transl. Med. **7**(11), 233 (Jun 2019)

26. Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. arXiv preprint arXiv:2404.16754 (2024)