Evaluating Medical Report Generation on Real Clinical Data: A Case Study

Dominik Hirsch[1[0009−0004−6335−150X]], Hannes Ulrich[2[0000−0002−8349−6798]], Heinz Handels[1,3[0000−0002−3499−4328]], and Jan Ehrhardt[1,3[0009−0004−6804−5587]]

[1] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany
dominik.hirsch@uni-luebeck.de
[2] Medical Informatics Laboratory, Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
[3] German Research Center for Artificial Intelligence (DFKI), Luebeck, Germany

Abstract. Automatic report generation for chest CT has advanced rapidly, yet how well pretrained models perform on real clinical data remains largely untested. We evaluate two published report generation models on an out-of-distribution clinical dataset and discuss the reproducibility challenges that frequently accompany freely available models. To enable the evaluation, we assemble and curate an institutional dataset of raw clinical data into model-ready inputs. Comparing it to a public benchmark reveals substantial differences, including more detailed localization, explicit progression descriptions, and frequent quantitative measurements in the clinical reports. Assessing the selected models with natural language generation and clinical efficacy metrics, we find that they only marginally outperform naive baselines. A hallucination analysis indicates that the evaluated models primarily reproduce training-set language while making limited use of the visual input. These findings underscore the importance of evaluating report generation models on realistic clinical data and of publishing complete reproduction information as a standard of scientific practice.

Keywords: clinical data · medical report generation · chest CT.

1 Motivation

Automating radiology report generation promises to relieve a significant documentation burden, and the release of the large-scale CT-RATE [1] dataset has rapidly accelerated model development for 3D chest CT. However, CT-RATE's curated nature does not reflect the noise and heterogeneity of routine hospital PACS data, raising the question of how well these models transfer to realistic clinical settings. Beyond generalization, a second concern is whether the models capture image-based findings, or merely reproduce learned linguistic patterns. A further, often overlooked barrier is reproducibility. Many published models are released incompletely, so that only few can be faithfully reproduced and evaluated on new data. We therefore evaluate the report generation models that we could reproduce on a curated real-world clinical chest CT dataset, providing an assessment of their applicability to clinical practice.

2 Materials and Methods

2.1 Related Work and Model Choice

Existing evaluations of chest CT report generation are dominated by curated public benchmarks. Where out-of-domain performance is examined, it is typically reported by the authors of a newly proposed model to demonstrate its generalization. Most such evaluations use a second public cohort [1,8], and while some do include private external institutional data [2], systematic third-party assessment of existing public models on private clinical data remains scarce. We therefore adopt the perspective of a clinician with an institutional dataset who wishes to assess how well existing report generation models perform on it.

Selecting models for this setting, we found reproducibility and availability to be the primary practical constraints, and report three observations from our experience across several repositories. First, code and weights are typically released months after acceptance and are often initially incomplete, so the publicly reproducible models tend to lag behind the latest reported approaches. Second, release alone does not guarantee reproducibility. Undocumented preprocessing, withheld evaluation scripts, convoluted code, or unpinned dependencies can prevent full reproduction or degrade results. Third, for large models, unreleased weights leave only retraining, so a reproduction gap persists that is driven purely by hardware availability. Released models should therefore be accompanied by all information needed to independently verify and build on them.

Accordingly, we restrict our attention to models with public code and weights, considering [1,3–10]. We exclude candidates whose training data are unavailable or whose training domain or label space differs too strongly from ours, leaving five models. Two proved reproducible enough for a fair comparison on our institutional data and form the basis of our evaluation. CT-CHAT [1] embeds CT volumes with CT-CLIP and fine-tunes a Llama-based language model on a large VQA dataset, enabling question answering and report generation. Reg2RG [10] additionally uses masks of selected organs to produce region-specific text, disentangling local and global visual information.

2.2 Data

Private Clinical Dataset We collected a clinical chest CT dataset from adult patients in the second largest hospital in Germany, exported from the hospital PACS using a clinically approved anonymization pipeline in compliance with data protection regulations. Informed consent was obtained from all patients. Reflecting routine practice, the raw data are heterogeneous, including scans of other anatomical regions, single-slice acquisitions, and scanned documents.

Using a locally hosted, privacy-preserving curation pipeline, we filter this data to diagnostic chest CT studies with reports and prepare model-ready inputs. Reports are translated to the downstream model's target language, and binary abnormality labels are extracted with a locally deployed Llama 3.1 70B [12] to form

pseudo-ground-truth for clinical efficacy evaluation. Volumes with unsuitable acquisition properties (large slice spacing, low in-plane resolution, single-slice acquisitions) or not depicting the lung region are removed, and the remainder are standardized to each model's input format via consistent orientation, resampling, and cropping. The final curated dataset comprises 1055 chest CT volumes with corresponding ground-truth reports.

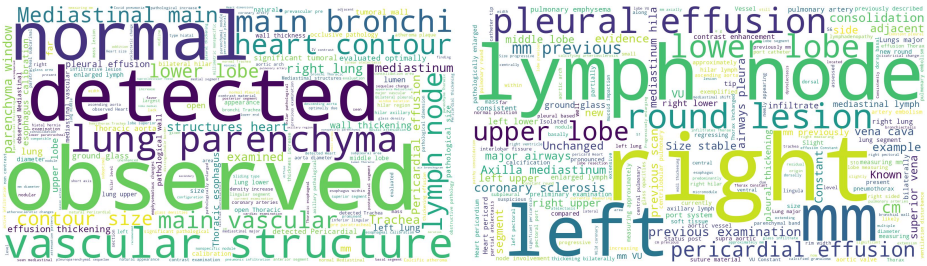
Public Datasets We compare our private dataset to the CT-RATE [1] validation set, which comprises over 1500 unique radiology reports paired with chest CT volumes and annotated with 18 binary anomaly labels. Since Reg2RG outputs region-level reports, we additionally use RadGenome-ChestCT [11], which extends CT-RATE with 10 regional masks per volume and corresponding region-level annotations. For the hallucination analysis, we use the LIDC-IDRI [13] lung nodule dataset with label masks from [14], comprising 882 samples with all nodules of maximum in-plane diameter ≥ 3 mm present in the masks.

2.3 Evaluation Setup

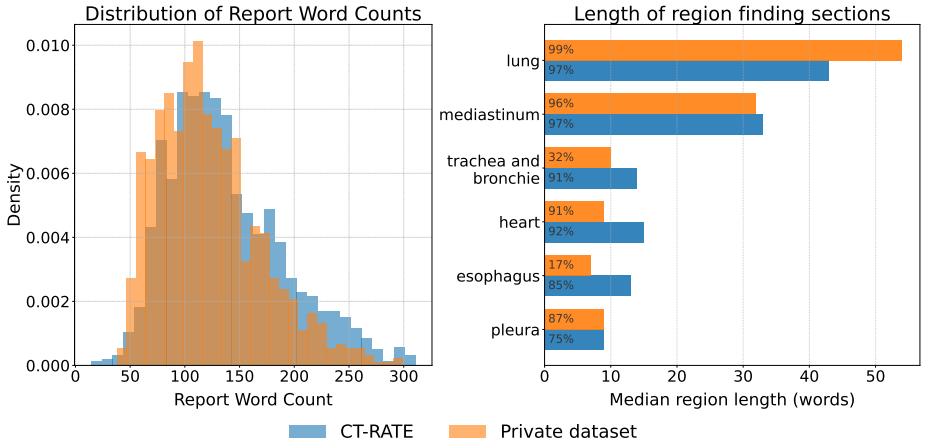
For CT-RATE and RadGenome-ChestCT, ground-truth labels and reports are provided. For our dataset, ground truth reports exist, while binary anomaly labels and region-level reports are derived using LLMs. We extract the same 18 binary anomaly labels as CT-RATE using Llama 3.3 70B, achieving an F1-score of 0.919 against the CT-RATE-provided validation labels. Since CT-RATE labels are also model-derived and therefore noisy, this level of agreement validates the use of the LLM-based procedure to generate binary anomaly labels for our clinical dataset. Likewise, Llama 3.1 70B is used to split our reports into region-level granularity to compare against the region-level reports produced by Reg2RG.

Model performance is assessed with clinical efficacy and natural language generation metrics. For clinical efficacy, binary labels extracted from the generated reports are compared to the model-derived ground truth labels of CT-RATE and our dataset using macro-averaged precision, recall, F1-score, and balanced accuracy over the 18 anomaly labels. CT-CHAT and Reg2RG are evaluated on identical ground-truth labels and are thus comparable. For language evaluation, we restrict to the Findings section, comparing CT-CHAT against raw reports and Reg2RG against the concatenation of region reports using BLEU-1, BLEU-4 [15], METEOR [16], and ROUGE-L [17].

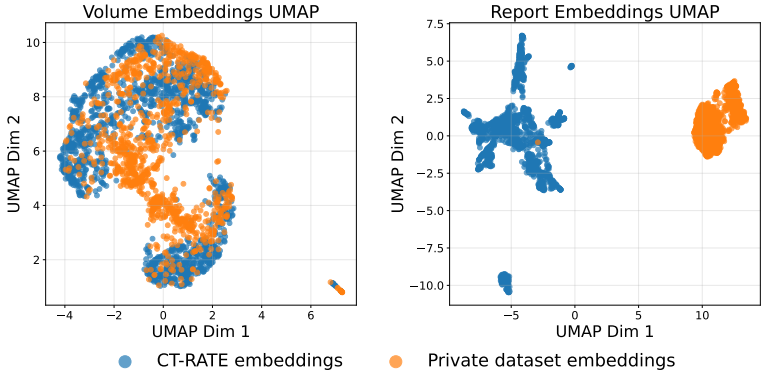
In the following sections, we conduct three analyses. In Section 3.1, we compare our dataset to CT-RATE and its extension RadGenome-ChestCT. In Section 3.2, we report metrics for the examined models and compare them against naive baselines that reproduce only the anomaly distribution or the linguistic patterns of the training data. In Section 3.3, we perform a hallucination analysis on LIDC-IDRI. We first evaluate how accurately CT-CHAT answers questions about the presence of lung nodules. Then, we generate reports for the LIDC volumes, extract the predicted nodules, and assess the correctness of the predictions.



(a) Word clouds illustrating frequent words in the datasets. Left: CT-RATE, Right: Private dataset.



(b) Distribution of report word counts (left) and median region report length of selected regions (right). Percentage values indicate the share of reports containing the region.



(c) UMAP projections of embeddings. Left: BTB3D volume embeddings (100 PCA components). Right: BioClinicalBERT report embeddings.

Fig. 1: Overview of dataset characteristics and embeddings. (a) Word clouds illustrating frequent words in the CT-RATE and private dataset, (b) Report lengths and region shares, (c) UMAP embeddings.

3 Results

3.1 Dataset Comparison

Figure 1 summarizes characteristics of the CT-RATE validation samples and the 1055 volume-report pairs in our dataset. The word clouds in 1a reveal distinct vocabulary patterns. CT-RATE mainly features general descriptors such as 'detected', 'observed', and 'normal', whereas the private dataset contains more specific, target-oriented terminology, including frequent positional references ('left', 'right', 'upper lobe', 'lower lobe'), temporal evolution ('previous', 'size stable', 'unchanged'), and measurements in 'mm'. 1b shows comparable report lengths across both datasets at full-report and region-specific levels, with lung and mediastinum comprising the largest share of text. Lung, mediastinum, heart, and pleura are the most consistently described regions, while trachea and bronchi and esophagus are mentioned more often in CT-RATE.

Fig. 1c presents UMAP projections of volume and report embeddings, extracted with BTB3D [8] and BioClinical-BERT [18], respectively. Despite partial 2D overlap, a k-nearest neighbor classifier on the top 100 PCA components separates the datasets well for volumes (accuracy 0.88, ROC-AUC 0.95) and perfectly for reports (1.0, 1.0). Overall, the datasets share similarities, as report length, referenced regions, and finding sets are largely comparable. However, the strong embedding separation reveals substantial differences in reporting style and imaging characteristics. The private dataset more frequently describes disease progression and contains more precise positional references, motivating an assessment of how well models trained on CT-RATE generalize to this cohort.

3.2 Quantitative results

Table 1 compares Reg2RG and two CT-CHAT configurations against three lower-bound baselines that operate without any volume input, isolating what can be achieved from training-set statistics and language patterns alone. *Freq. CT-RT* replicates the anomaly statistics of the CT-RATE training set, sampling labels from its frequency distribution and randomly matching them to evaluation samples (averaged over 1000 matchings). *Rep. CT-RT* and *Llama CT-RT* instead replicate training-set language. The former randomly matches real CT-RATE training reports to evaluation samples, while the latter matches volume-independent reports generated by a Llama 3.1 model fine-tuned on the CT-RATE training set (both averaged over 100 matchings). A model that has learned to use visual information should exceed these baselines.

On the CT-RATE evaluation set, all three models outperform the baselines on the NLG metrics and show markedly higher precision. However, this comes at the cost of recall, so their F1 does not exceed the naive label-frequency baseline. The clearest evidence that they use visual information is balanced accuracy, which rises a few points above chance. On the private dataset, the NLG metrics collapse toward the baseline level, revealing that improved CT-RATE scores stemmed largely from reproducing training-set report language, which does not

Table 1: NLG and CE metrics of the examined models compared to baseline approaches. All values are percentages. *Reg2RG produces region-level reports evaluated against the concatenated RadGenome-ChestCT region reports, whereas all other models are evaluated against the raw CT-RATE reports.

Data	Model	BLEU-1	BLEU-4	RG-L	MTR	Precision	Recall	F1	BalAcc
CT-RT	Freq. CT-RT	-	-	-	-	19.00	18.90	18.93	49.98
	Llama CT-RT	43.30	22.18	36.42	40.74	18.89	7.96	10.67	49.98
	Rep. CT-RT	42.35	19.13	32.23	39.73	18.88	18.46	18.64	49.91
	CT-CHAT 8B	46.62	25.55	38.83	43.62	37.14	16.07	17.16	53.66
	CT-CHAT 70B	46.31	24.59	37.85	43.12	37.85	14.75	16.56	53.69
	Reg2RG*	47.57	25.02	36.72	44.28	39.64	13.69	18.65	53.81
Own	Freq. CT-RT	-	-	-	-	25.53	18.96	19.94	50.01
	Llama CT-RT	15.93	2.04	12.47	14.43	25.54	8.01	10.71	50.00
	Rep. CT-RT	19.85	2.46	13.24	16.50	25.36	18.61	19.63	49.97
	CT-CHAT 8B	18.90	2.75	13.48	15.83	24.47	21.59	17.27	52.00
	CT-CHAT 70B	18.34	2.59	13.20	15.39	26.56	19.20	17.63	52.13
	Reg2RG*	22.27	2.72	13.67	17.58	30.67	14.35	17.44	51.85

Table 2: Hallucination analysis for CT-CHAT 8B. For each prompt we report how often the answer is "yes" in the training data (%Yes, Train) and how often the model answers "yes" at inference on LIDC (%Yes, Eval) (6/32 prompts shown).

Question	Train		Eval
	Samples	%Yes	%Yes
Are there any nodules in the lungs?	759	99	100
Are there any masses or nodules in the lungs?	98	0	0
Are there any nodules in the lung parenchyma?	77	86	100
Are there any masses or nodules in the lung parenchyma?	138	0	0
Can you see any nodules in the lungs?	77	100	100
Are there any nodular or infiltrative lesions in the lung parenchyma?	613	0	0

transfer to a different reporting style. Clinical efficacy degrades in parallel, and balanced accuracy retains only a marginal edge over chance. The models thus exhibit a weak but genuine visual signal overwhelmed by learned language priors, and scaling CT-CHAT from 8B to 70B yields no improvement.

3.3 Hallucination Analysis

We investigate hallucinations using the 882 LIDC-IDRI CT volumes with annotated lung nodules (>3 mm). Lung lobe segmentations are used to derive the anatomical location of each nodule and establish ground truth.

We evaluate CT-CHAT 8B using binary questions about nodule presence in the lungs and individual lobes. Across different prompts and answer modes, the model achieves balanced accuracies close to random chance (maximum 0.517).

Table 3: Lung nodule prediction results on LIDC. Volume-level rows count how many of the 882 nodule-bearing volumes yield a report mentioning a nodule. Lobe-level rows then narrow these reports step by step: mentions that specify laterality, that further specify a lobe, and that further state a size $\geq$3 mm. *Correct* are those whose stated lobe matches the ground truth (1743 nodule-bearing lobes total).

Volume-level (**882 CT volumes**, each containing $\geq$1 nodule)			
	CT-CHAT 8B	CT-CHAT 70B	Reg2RG
No mention / explicit absence of nodules	364	368	635
Reports mentioning nodule(s)	518	514	247
Lobe-level (reference: **1743 nodule-bearing lobes**)			
Predicted nodules	518	514	303
+ laterality (L/R)	1	0	168
+ specific lobe	1	0	160
+ size $\geq$ 3 mm	1	0	104
Correct (lobe = GT)	1	0	26

We observe that predictions are largely independent of the visual input. To investigate this behavior, we analyze 32 frequently occurring yes/no questions from the training data (Table 2). In 31/32 cases, training answers exhibit a class imbalance above 90%, and the model consistently predicts the majority training label. Similar behavior is observed for CT-CHAT 70B, suggesting that the models primarily reproduce textual priors rather than visual evidence.

We further evaluate generated reports from all models (Table 3). CT-CHAT frequently generates the generic statement 'There are millimetric nonspecific nodules in both lungs' without providing verifiable locations or sizes. Reg2RG produces more detailed nodule descriptions, but many are not supported by ground truth annotations. The results highlight that the examined report generation models are prone to hallucinations when applied to clinical data.

4 Summary

Taking the perspective of a clinician evaluating report generation models on institutional data, we assessed how well publicly available models trained on CT-RATE generalize to a real clinical cohort. Reproducibility proved a central obstacle. Delayed or incomplete releases, missing preprocessing and evaluation details, and unpublished weights left only CT-CHAT and Reg2RG reproducible enough to evaluate. We curated a private dataset of 1055 chest CT volumes with reports and LLM-derived pseudo-ground-truth labels, which exhibits similarities with public benchmarks but offers more precise localization, explicit progression, and frequent measurements. Across both datasets, the models only marginally outperformed naive baselines, and on the private cohort their language metrics collapsed to baseline level. A hallucination analysis traced this to the models

reproducing training-set language while largely ignoring the visual input, underscoring the need to evaluate report generation on realistic clinical data.

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

References

1. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nat. Biomed. Eng. (2026). https://doi.org/10.1038/s41551-025-01599-y
2. Wang, Z., Chen, F., Yu, C., Li, Z., Zheng, Y., Wang, J., et al.: Astra: A Generalizable Report Generation Foundation Model for 3D Computed Tomography. arXiv preprint arXiv:2605.31437 (2026)
3. 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 (2025). https://doi.org/10.1038/s41467-025-62385-7
4. 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:2404.00578 (2024).
5. Deng, X., He, X., Bao, J., Zhou, Y., Cai, S., Cai, C., et al.: MvKeTR: Chest CT Report Generation With Multi-View Perception and Knowledge Enhancement. IEEE J. Biomed. Health Inform. **29**(11), 8279–8292 (2025). https://doi.org/10.1109/JBHI.2025.3581289
6. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-LLaMA: Towards Large Language Model-Driven CT Report Generation. In: MICCAI 2025, pp. 141–151 (2026). https://doi.org/10.1007/978-3-032-04981-0_14
7. Shi, Y., Zhu, X., Wang, K., Hu, Y., Guo, C., Li, M., et al.: Med-2E3: A 2D-Enhanced 3D Medical Multimodal Large Language Model. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2754–2759. IEEE (2025). https://doi.org/10.1109/BIBM66473.2025.11357048
8. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Yang, D., Guo, P., et al.: Better Tokens for Better 3D: Advancing Vision-Language Modeling in 3D Medical Imaging. arXiv preprint arXiv:2510.20639 (2025)
9. Liang, R., Ma, Y., Xing, Y., Fan, Z., Pan, J., Sun, C., et al.: Beyond the Embedding Bottleneck: Adaptive Retrieval-Augmented 3D CT Report Generation. arXiv:2603.15822 (2026).
10. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large Language Model With Region-Guided Referring and Grounding for CT Report Generation. IEEE Trans. Med. Imaging **44**(8), 3139–3150 (2025). https://doi.org/10.1109/TMI.2025.3559923
11. Zhang, X., Wu, C., Zhao, Z., Lei, J., Tian, W., Zhang, Y., et al.: Development of a large-scale grounded vision language dataset for chest CT analysis. Sci. Data **12**(1), 1636 (2025). https://doi.org/10.1038/s41597-025-05922-9
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al.: The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783 (2024)
13. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., et al.: Data from LIDC-IDRI [Data]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX
14. Fedorov, A., Hancock, M., Clunie, D., Brockhausen, M., Bona, J., Kirby, J., et al.: Standardized representation of the TCIA LIDC-IDRI annotations using DICOM. The Cancer Imaging Archive (2018). https://doi.org/10.7937/TCIA.2018.h7umfurq

15. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proc. 40th Annu. Meeting Assoc. Comput. Linguist. (ACL '02), pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135
16. Lavie, A., Agarwal, A.: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proc. 2nd Workshop Stat. Mach. Transl., pp. 228–231 (2007).
17. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out, pp. 74–81 (2004).
18. Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., et al.: Publicly Available Clinical BERT Embeddings. In: Proc. 2nd Clinical Natural Language Processing Workshop, pp. 72–78 (2019). https://doi.org/10.18653/v1/W19-1909