Balancing Retrieved Evidence for 3D CT Report Generation

Yue Heng[1], Yuichiro Hayashi[1], Masahiro Oda[2,1], and Kensaku Mori[1,2,3]

[1] Graduate School of Informatics, Nagoya University, Furou-cho, Chikusa-ku, Nagoya, 464-8601, Aichi, Japan
[2] Information Technology Center, Nagoya University, Furou-cho, Chikusa-ku, Nagoya, 464-8601, Aichi, Japan
[3] Research Center for Medical Bigdata, National Institute of Informatics, Hitotsubashi, Chiyoda-ku, 101-8430, Tokyo, Japan

Abstract. Automatic radiology report generation from 3D computed tomography (CT) remains challenging because generated reports may contain unsupported findings or omit clinically relevant abnormalities. We propose an evidence-aware retrieval-augmented framework that retrieves similar training cases with a 3D CT–text contrastive model and converts retrieved reports into compact positive and negative finding checklists. These checklists are presented as auxiliary evidence to be verified against the current CT volume rather than as text to copy. We further apply RAG-aware supervised fine-tuning so that the generator learns to use this evidence during report generation. Evaluation on the CT-RATE validation set using the official RadBERT-based 18-label clinical efficacy protocol shows that prompt-only RAG consistently reduces hallucinated labels but increases omissions, resulting in significant CE macro-F1 degradation across all four retrieval strategies (all $p < 0.001$). The RAG-aware fine-tuned variants show partial numerical recovery from this degradation. Fusion-based SFT shows no statistically detectable difference in CE macro- or micro-F1 from the no-RAG baseline while significantly reducing hallucinated labels, whereas I2I achieves the highest observed BLEU and RadGraph ER F1 but retains a CE-F1 gap. Overall, the results reveal an over-conservative failure mode of prompt-only retrieval conditioning and highlight the need to jointly evaluate hallucination, omission, and clinical efficacy when incorporating retrieved evidence into 3D CT report generation.

Keywords: 3D CT report generation · Retrieval-augmented generation · Radiology report generation · Evidence-aware fine-tuning · Vision-language model

1 Introduction

Radiology report generation aims to automatically describe clinically relevant findings from medical images, reducing radiologists' workload and improving reporting consistency. While this task has been widely studied for 2D chest radiographs [3, 12], extending it to 3D computed tomography (CT) is substantially

more challenging: a single chest CT study contains hundreds of slices, complex anatomical structures, and spatially distributed findings, requiring models to identify clinically meaningful abnormalities across volumetric data rather than a single 2D projection.

Recent progress in 3D CT vision–language modeling has made volumetric report generation increasingly feasible. Large-scale paired 3D CT and report data, together with generalist 3D CT vision–language models such as CT-CHAT [5], have supported contrastive representation learning and instruction-following generation. Dedicated report generation methods such as CT2Rep [4] and CT-GRAPH [7] further improve volumetric feature aggregation and anatomy-aware reasoning. Nevertheless, generated reports may still contain clinically unsupported findings or omit relevant abnormalities.

Retrieval-augmented generation (RAG) [8] offers a promising direction for improving factuality by providing similar prior cases as external clinical context. Recent medical multimodal RAG methods have emphasized that retrieved context must be carefully selected and calibrated. RULE [14] studies reliable context utilization and reduces over-reliance on potentially misleading retrieved evidence, whereas MMed-RAG [13] combines domain-aware retrieval and adaptive context selection to improve factuality across medical vision–language tasks. These studies demonstrate the potential of retrieval, but also show that irrelevant or excessive context can adversely affect generation. In 3D CT report generation, retrieved cases may contain abnormalities absent from the current CT, while relevant findings in the query may be missing from the retrieved reports. Simply concatenating retrieved reports to the input therefore does not guarantee faithful generation: the generator may ignore the evidence, copy unsupported findings, or become overly conservative and omit true findings.

The closest recent work is AdaRAG-CT [9], which uses a Self-RAG [1]-inspired mechanism to determine *when* retrieved context should be incorporated during generation. Our work addresses a complementary question: once similar cases are retrieved, *how* should their information be represented? We convert retrieved reports into structured positive and negative finding candidates that are presented as auxiliary evidence to be verified against the current CT volume.

To this end, we retrieve similar training cases using image-to-image (I2I), image-to-text (I2T), fusion, and overlap-aware consensus retrieval. Retrieved reports are converted into compact positive and negative finding checklists, which are presented as auxiliary evidence to be verified against the current CT volume. We further perform RAG-aware supervised fine-tuning using the same evidence-conditioned prompt format at training and inference time.

Using the official CT-RATE [5] RadBERT-based [15] 18-label clinical efficacy (CE) protocol, prompt-only RAG consistently reduced hallucinated labels but increased omissions, resulting in lower CE macro-F1 across all four retrieval strategies. RAG-aware SFT provided partial numerical recovery, although no RAG configuration exceeded no-RAG in CE macro-F1. These results reveal an over-conservative failure mode and emphasize the need to evaluate both hallucination and omission.

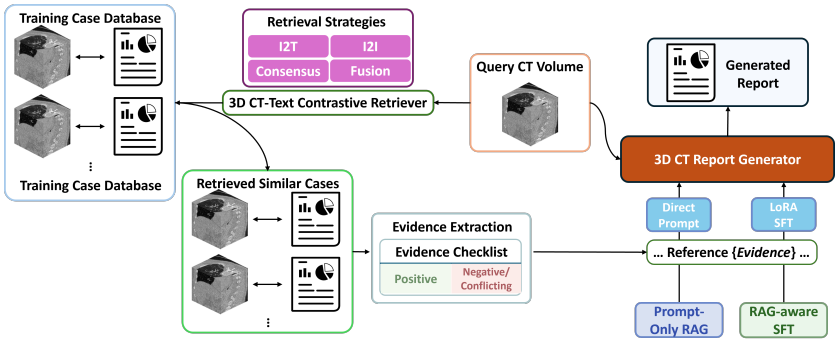

Fig. 1. Overview of the proposed evidence-aware retrieval-augmented framework for 3D CT report generation. A 3D CT-text contrastive retriever retrieves similar training cases using multiple retrieval strategies. Retrieved reports are converted into a structured finding-level evidence checklist, which is used as auxiliary evidence for prompt-only RAG or RAG-aware supervised fine-tuning. The generator produces the final report conditioned on both the current CT volume and the retrieved evidence.

Our main contributions are:

- An evidence-aware 3D CT report-generation framework comparing four retrieval strategies.
- A structured checklist that converts retrieved reports into positive, negative, and conflicting finding evidence.
- A paired-bootstrap analysis of prompt-only and RAG-aware SFT, revealing an over-conservative failure mode and different observed hallucination–omission profiles.

2 Method

2.1 Overview

We propose an evidence-aware retrieval-augmented framework for 3D CT report generation. Given a query CT volume, our method retrieves similar cases from the training set using a 3D CT-text contrastive retriever, converts the retrieved reports into a compact finding-level evidence checklist of positive and negative candidate findings, and conditions a 3D CT report generator on both the CT volume and this checklist. Unlike directly concatenating retrieved reports, we treat retrieved cases as structured candidate evidence rather than ground truth: the generator is instructed to verify each candidate finding against the current CT volume. Figure 1 illustrates the overall pipeline.

2.2 3D CT Report Generator

The baseline report generator consists of an 3D ResNet-based visual encoder, a Q-Former visual aggregator, and a Qwen2.5-7B-Instruct language model. The

visual encoder extracts volumetric features from the input CT volume. The Q-Former aggregates these features into 64 visual query tokens, which are projected into the language-model embedding space and provided as visual context to the decoder.

The generator is trained with an instruction-based report-generation prompt. The training objective is the autoregressive language-modeling loss computed only over the target report tokens, while the instruction and visual-context tokens are excluded from the loss.

2.3 Retrieval and Evidence Construction

Retrieval strategies. The retrieval database was restricted to the training split. Image and text embeddings were ℓ_2-normalized, and retrieval similarities were computed using cosine similarity. To prevent self-retrieval and patient-level information leakage, the query volume itself and all database volumes belonging to the same patient were excluded from the candidate pool.

For image-to-image retrieval (I2I), the query CT image embedding was compared with the image embeddings in the retrieval database. For image-to-text retrieval (I2T), the query image embedding was compared with the database report embeddings. In both cases, the five highest-scoring cases were retained.

Fusion retrieval combined standardized I2I and I2T similarity scores with equal weights:

$$\mathbf{s}_{\text{fusion}} = 0.5\, z(\mathbf{s}_{\text{I2I}}) + 0.5\, z(\mathbf{s}_{\text{I2T}}), \tag{1}$$

where $z(\cdot)$ denotes score standardization over the retrieval database.

For overlap-aware consensus retrieval, we first selected the top 50 candidates independently using I2I and I2T retrieval and formed the union of the two candidate sets. Candidates appearing in both top-50 lists received an additional overlap bonus:

$$\mathbf{s}_{\text{consensus}} = 0.4\, z(\mathbf{s}_{\text{I2I}}) + 0.4\, z(\mathbf{s}_{\text{I2T}}) + 0.2\, \mathbf{I}_{\text{overlap}}, \tag{2}$$

where $\mathbf{I}_{\text{overlap}}$ is a binary indicator vector whose element equals 1 when the corresponding candidate appears in both top-50 lists and 0 otherwise.

Rule-based evidence extraction. We used a deterministic extractor covering 27 manually defined finding concepts and their lexical variants. Reports were sentence-segmented, and positive or negative mentions were identified using regular expressions and predefined negation cues within a 180-character window. For each query, findings were aggregated across the top five retrieved reports. Findings supported positively by at least two cases without negative mentions were treated as high-confidence positive evidence; findings with both positive and negative support were marked as conflicting evidence; and findings negated in at least two cases without positive support were treated as negative evidence. Low-confidence and four non-target findings were excluded, and the checklist was limited to 12 items. Conflicting findings were retained explicitly as conflicting candidate evidence rather than resolved through majority voting, and the generator was instructed to verify all checklist items against the current CT volume.

2.4 Prompt-Only RAG and RAG-Aware Fine-Tuning

We evaluate two approaches for incorporating retrieved evidence. In prompt-only RAG, the baseline generator remains fixed and receives the evidence checklist only at inference time. This setting evaluates whether evidence introduced without evidence-aware training can improve report generation.

To reduce the resulting train–inference mismatch, we perform RAG-aware supervised fine-tuning. For each training CT volume, patient-disjoint similar cases are retrieved using the protocol described in Section 2.3, and their reports are converted into an evidence checklist. The generator is then trained to produce the ground-truth report conditioned on both the current CT volume and the retrieved evidence.

During fine-tuning, the visual encoder is frozen, whereas the Q-Former is optimized. The language model is adapted using LoRA on the query, key, value, and output projection layers. We train four separate RAG-aware models, one for each retrieval strategy: I2I, I2T, fusion, and overlap-aware consensus retrieval.

3 Experiments and Results

3.1 Experimental Setup

We evaluate all methods on the CT-RATE [5] validation split, using only the training split as the retrieval database. We compare no-RAG, four prompt-only RAG variants, and four separately trained RAG-aware SFT variants using I2I, I2T, fusion, and overlap-aware consensus retrieval. All methods were evaluated on the same validation cases for paired comparison.

For RAG-aware SFT, Qwen2.5-7B-Instruct was adapted using LoRA on `q_proj`, `k_proj`, `v_proj`, and `o_proj`, with $r = 16$, $\alpha = 32$, and dropout 0.05. Each model was trained for 10 epochs with a learning rate of 2×10^{-4}, batch size 8, warm-up ratio 0.05, no weight decay, and bfloat16 precision. The checkpoint with the lowest internal-validation loss was used for final evaluation.

3.2 Evaluation Metrics

We evaluate generated reports with both text similarity and clinically motivated factuality metrics. Text similarity is measured by BLEU [11], ROUGE-L [10], and METEOR [2]; clinical entity/relation similarity by RadGraph ER F1 (RG-ER) [6]. For clinical efficacy, we adopt the official CT-RATE protocol: the RadBERT-based 18-label classifier is applied to generated reports, and the predicted labels are compared with the official CT-RATE validation labels to compute macro- and micro-averaged F1. We report macro- and micro-averaged F1 as our primary clinical efficacy metrics, together with the mean number of hallucinated findings (false positives) and omitted findings (false negatives) per case. The RadBERT classifier and the official validation labels were used only for evaluation and were not involved in retrieval, evidence extraction, checklist construction, or report generation.

Table 1. Results on the CT-RATE validation set. Macro/Micro F1, hallucination, and omission use the official RadBERT 18-label clinical efficacy protocol; BLEU/ROUGE-L/METEOR/RG-ER are auxiliary text-similarity metrics reported for reference. Hallucination/omission are mean counts per case. Bold indicates the best value within each group (Prompt-only / RAG-aware SFT).

Method	Macro F1↑	Micro F1↑	Hall.↓	Omit.↓	BLEU	ROUGE-L	METEOR	RG-ER
No-RAG	0.363	0.461	1.047	2.112	20.98	0.353	0.347	0.272
Prompt-only, Consensus	0.340	0.445	0.889	2.221	22.06	0.367	0.354	0.277
Prompt-only, Fusion	0.344	0.449	**0.886**	2.207	22.20	0.368	0.356	0.278
Prompt-only, I2I	0.343	0.445	**0.886**	2.222	22.33	0.369	0.357	0.279
Prompt-only, I2T	**0.344**	**0.451**	0.906	**2.193**	22.05	0.366	0.355	0.278
SFT, Consensus	**0.360**	0.457	1.074	**2.121**	21.41	0.369	0.350	0.270
SFT, Fusion	0.357	**0.458**	0.984	2.144	22.87	0.388	0.361	0.281
SFT, I2I	0.348	0.449	**0.883**	2.207	**22.98**	**0.388**	**0.361**	**0.284**
SFT, I2T	0.355	0.451	0.964	2.175	22.25	0.382	0.356	0.276

3.3 Main Results

Table 1 shows that prompt-only RAG consistently reduced hallucinated labels but increased omissions and lowered CE macro-F1 across all four retrieval strategies. For example, fusion reduced hallucination from 1.047 to 0.886 labels per case, while macro-F1 decreased from 0.363 to 0.344 and omission increased from 2.112 to 2.207. This pattern indicates a more conservative rather than more clinically effective generator.

RAG-aware SFT numerically recovered part of this degradation, although no variant exceeded no-RAG in CE macro-F1. Fusion SFT showed a favorable balance between CE F1, hallucination, and omission. I2I achieved the highest BLEU and RG-ER and tied with Fusion on ROUGE-L and METEOR, but had lower CE macro-F1 and more omissions. Because strategies were not directly compared statistically, their numerical differences are interpreted descriptively.

3.4 Statistical Analysis

We performed case-level paired bootstrap resampling ($B = 2{,}000$) over validation cases shared by all methods. Macro-F1, hallucination, and omission were recomputed for each resample, and paired differences relative to no-RAG were obtained with two-sided p-values.

As shown in Table 2, all prompt-only variants had lower macro-F1, fewer hallucinations, and more omissions than no-RAG, with confidence intervals excluding zero. Among the SFT variants, Fusion showed a small macro-F1 difference and fewer hallucinations, while its omission interval included zero. I2I produced the largest hallucination reduction but also lower macro-F1 and more omissions. All comparisons were made against no-RAG; they do not establish direct superiority between retrieval strategies.

Table 2. Paired bootstrap significance analysis (B=2,000 resamples) for all eight RAG variants relative to the no-RAG baseline. Values are mean paired differences with 95% confidence intervals. Positive differences are better for macro F1, while negative differences are better for hallucination and omission. Bold p-values indicate significance at $\alpha = 0.05$.

Method (vs. no-RAG)	Δ Macro F1	Δ Hall.	Δ Omit.
Prompt-only RAG			
Consensus	-0.0228 [-0.0321, -0.0134], < 0.001	-0.1586 [-0.1872, -0.1269], < 0.001	+0.1093 [0.0819, 0.1373], < 0.001
Fusion	-0.0187 [-0.0274, -0.0100], < 0.001	-0.1612 [-0.1925, -0.1299], < 0.001	+0.0949 [0.0680, 0.1219], < 0.001
I2I	-0.0199 [-0.0295, -0.0105], < 0.001	-0.1616 [-0.1925, -0.1306], < 0.001	+0.1099 [0.0826, 0.1369], < 0.001
I2T	-0.0188 [-0.0282, -0.0097], < 0.001	-0.1416 [-0.1716, -0.1113], < 0.001	+0.0813 [0.0530, 0.1086], < 0.001
RAG-aware SFT			
Consensus	-0.0032 [-0.0145, 0.0074], 0.549	+0.0263 [-0.0113, 0.0640], 0.157	+0.0097 [-0.0223, 0.0420], 0.543
Fusion	-0.0056 [-0.0171, 0.0046], 0.277	-0.0636 [-0.0996, -0.0256], 0.001	+0.0320 [-0.0010, 0.0650], 0.059
I2I	-0.0148 [-0.0262, -0.0045], 0.007	-0.1646 [-0.2002, -0.1269], < 0.001	+0.0949 [0.0640, 0.1279], < 0.001
I2T	-0.0079 [-0.0197, 0.0034], 0.150	-0.0829 [-0.1226, -0.0430], < 0.001	+0.0633 [0.0303, 0.0963], < 0.001

4 Discussion and Conclusion

This study investigates how retrieved evidence should be represented and used in 3D CT report generation. Across all four retrieval strategies, prompt-only RAG reduced hallucinated labels but increased omissions, resulting in lower CE macro-F1 than the no-RAG baseline. This consistent error-profile shift indicates that introducing retrieved evidence only at inference time can make the generator overly conservative rather than improving overall clinical efficacy.

The RAG-aware SFT variants showed partial numerical recovery from the degradation observed under prompt-only conditioning, although none exceeded the no-RAG baseline in CE macro-F1. Fusion SFT showed a favorable observed trade-off, with CE macro- and micro-F1 close to no-RAG, fewer hallucinated labels, and a relatively small increase in omissions. Consensus SFT produced the highest macro-F1 among the SFT variants but did not reduce hallucination relative to no-RAG. I2T SFT reduced hallucinated labels while increasing omissions, whereas I2I SFT achieved the highest observed BLEU and RadGraph ER F1 but showed lower CE macro-F1 and greater omission.

One possible explanation is that retrieved negative evidence may suppress related true findings during generation. Differences in the composition of the retrieved evidence may also contribute to the observed error profiles across retrieval strategies. However, these mechanisms were not directly evaluated, and the numerical differences between strategies should be interpreted descriptively rather than as evidence of direct strategy-level superiority.

The results also illustrate a discrepancy between text-based similarity and standardized clinical efficacy. The retrieval configuration with the highest BLEU or RadGraph ER F1 did not achieve the highest CE F1, suggesting that improvements in lexical or relation-level overlap do not necessarily correspond to improved disease-label coverage. This finding highlights the importance of evaluating both hallucination and omission when applying retrieval augmentation to medical report generation.

Our work is complementary to adaptive retrieval methods that study when retrieved context should be introduced. We instead focus on how retrieved re-

ports are represented after retrieval and show that finding-level evidence conditioning can systematically alter the hallucination–omission trade-off. The proposed checklist framework therefore provides a structured setting for analyzing how retrieved clinical evidence affects 3D CT report generation.

This study has several limitations. First, clinical efficacy was evaluated using the official CT-RATE 18-label RadBERT classifier, which covers a fixed set of abnormalities and does not capture all clinically relevant report errors. Second, the rule-based evidence extractor was not independently validated and does not explicitly model epistemic uncertainty, anatomical scope, or temporal context. The observed behavior may therefore reflect both retrieval noise and evidence-extraction errors. Third, the statistical analysis primarily compared each RAG configuration with no-RAG rather than directly comparing retrieval strategies. Fourth, because no matched continued-training control without retrieved evidence was included, the effect of evidence conditioning cannot be completely separated from that of additional fine-tuning. Finally, evaluation was limited to the CT-RATE validation split from a single data source, without external or radiologist evaluation.

Future work will investigate more robust evidence extraction, controlled analyses of positive and negative evidence, adaptive retrieval calibration, external validation, and expert assessment of the observed omission patterns.

In conclusion, prompt-only checklist conditioning produced a consistent over-conservative failure mode in 3D CT report generation, reducing hallucinated labels at the cost of increased omissions and lower CE macro-F1. RAG-aware SFT partially recovered this degradation numerically, but no evaluated RAG configuration exceeded no-RAG in CE macro-F1. These findings emphasize that retrieved evidence must be carefully represented, calibrated, and evaluated across multiple clinical error dimensions.

Acknowledgments. This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) SIP2023E23, JSPS KAKENHI 24H00720 and JST CREST (JPMJCR20D5).

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

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: The Twelfth International Conference on Learning Representations (2024)
2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449 (2020)

4. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated radiology report generation for 3D medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
5. 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.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y
6. Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M., Ng, A., Langlotz, C., Rajpurkar, P.: RadGraph: Extracting clinical entities and relations from radiology reports. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran (2021)
7. Kalisch, H., Hörst, F., Kleesiek, J., Herrmann, K., Seibold, C.: CT-GRAPH: Hierarchical graph attention network for anatomy-guided CT report generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6834–6843 (2025)
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
9. Liang, R., Ma, Y., Xing, Y., Fan, Z., Pan, J., Sun, C., Li, L., Gong, K., Xu, J.: Beyond the embedding bottleneck: Adaptive retrieval-augmented 3D CT report generation. arXiv preprint arXiv:2603.15822 (2026)
10. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
12. Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., et al.: Collaboration between clinicians and vision–language models in radiology report generation. Nature Medicine **31**(2), 599–608 (2025)
13. Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., Yao, H.: MMed-RAG: Versatile multimodal RAG system for medical vision language models. In: The Thirteenth International Conference on Learning Representations (2025)
14. Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., Yao, H.: RULE: Reliable multimodal RAG for factuality in medical vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1093. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.emnlp-main.62
15. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), e210258 (2022)