An Adaptive Framework for Evaluating Radiology Reports with Limited Annotated Data

Vera Novikova[0009−0002−3288−6646], Yakov Pchelintsev[0000−0002−2017−5866], Alexander Khvostikov[0000−0002−4217−7141], and Andrey Krylov[0000−0001−9910−4501]

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia
nov.vera27@gmail.com, pchelintsev@cs.msu.ru, khvostikov@cs.msu.ru, kryl@cs.msu.ru

Abstract. Evaluating automatic radiology report generation is challenging due to the strict requirement for clinical correctness. While specialized evaluation metrics have recently emerged, they are predominantly designed for English, limiting their direct application to other languages without retraining or validating the corresponding models on annotated datasets of sufficient size. In this paper, we introduce an adaptive framework for Russian-language chest X-ray reports evaluation under limited annotated data. The proposed method combines three complementary component scores: interpretable clinical consistency analysis, text similarity via translation, and multimodal image-report alignment. The framework automatically selects the optimal configuration of component metrics and estimates fusion weights using nested cross-validation. We validate our approach against translated expert annotations derived from the RadEvalX dataset. Our results show that clinically aligned metrics for non-English medical reports can be successfully built by adapting and fusing pretrained textual and vision-language models. Finally, we analyze the shortcomings of the proposed approach to provide suggestions for future improvements.

Keywords: Radiology report evaluation · Evaluation metrics · Chest X-ray · Clinical natural language processing · Vision-language models · Multilingual medical natural language processing.

1 Introduction

Automatic chest X-ray report generation is increasingly studied as a tool for reducing routine workload and improving consistency in radiological documentation [15]. However, progress in this area depends not only on better generation models, but also on reliable evaluation. A generated report can be close to a reference report at the lexical level while changing a negation, omitting a finding, or placing an abnormality in the wrong anatomical region [16,20]. Such errors are clinically relevant but may receive acceptable scores from general-purpose natural language generation metrics.

For English radiology reports, a number of specialized evaluators have been developed that move beyond surface text similarity [11,12,13,16,19,20,21,23]. These include metrics based on clinical labels and structured report graphs, learned combinations of automatic scores, models that predict radiologist error counts, and recent LLM-based evaluators. Although such methods have improved agreement with expert assessment, they are not directly applicable to other languages. The development of a dedicated Russian evaluator is constrained by the limited availability of expert-annotated radiology datasets. Consequently, adapting and combining existing evaluation methods offers a more practical approach to constructing a metric for evaluating the quality of automatically generated Russian radiology reports.

We address the following problem. Given a chest X-ray image, a reference report, and a candidate report in Russian, the goal is to compute an automatic score that aligns with expert annotation. We use existing components as candidate features, then select and weight them without requiring extensive training data. The framework combines a Russian clinical component, translated English report-comparison metrics, and image-text consistency scores. The component choice and fusion weights are optimized inside a nested cross-validation procedure.

The contributions of this work are as follows:

1. We construct a Russian-language evaluation setting for chest X-ray reports by aligning translated RadEvalX report pairs with IU X-Ray images and expert error counts.
2. We propose an adaptive framework that combines Russian clinical consistency, translated reference-based textual metrics, and image-report alignment scores.
3. We evaluate the complete component-selection and fusion procedure with nested cross-validation to reduce optimistic bias under limited annotation.

2 Related Work

Over the past several decades, a diverse array of metrics has emerged to evaluate automatically generated text. Lexical and semantic text generation metrics assess quality by measuring word or token similarity between candidate and reference reports. Metrics like BLEU [17] and ROUGE [13] evaluate matching n-gram sequences or longest common subsequences. METEOR [1] incorporates word stemming and synonym matching. More recent embedding-based metrics such as BERTScore [21] compare contextual token representations to calculate similarity matrices. These approaches are computationally efficient but remain insensitive to critical medical modifications.

Specialized clinical metrics extract structured representations of medical content to incorporate domain knowledge. F1CheXbert represents reports as binary pathology vectors using the CheXbert labeler [11,19]. F1RadGraph [12] extracts clinical entities and relations as a graph, measuring overlap between candidate and reference structures. RadCliQ [20] fits a regression model over existing lexical

and clinical metrics to predict error count. Similarly, CREPE [5] uses a biomedical encoder and multiple regression heads to predict error frequencies across categories like false positives, omissions, or incorrect locations. RaTEScore [23] employs a specialized named entity recognition model and medical embeddings to align entity pairs based on a type penalty matrix.

Generative language models allow interpretable evaluation. The GREEN metric utilizes a language model to output structured error notations covering omissions, mislocalizations, and severity errors, alongside a normalized clinical score [16]. FineRadScore [10] frames evaluation through a minimum edit distance paradigm where a language model identifies string-level modifications, assigns a severity level, and provides text feedback. GEMA-Score [22] utilizes a multi-agent architecture where separate agents evaluate specific clinical dimensions, completeness, and readability. These generative metrics offer explanatory feedback but require substantial computational resources.

While the aforementioned evaluation frameworks are primarily optimized for the English language, adapting automated metrics to the Russian medical domain is severely constrained by the scarcity of annotated clinical data. To overcome this limitation, we introduce an adaptive approach to analyzing Russian chest X-ray reports, which, to the best of our knowledge, is the first specialized evaluation framework tailored to Russian radiology reports.

3 Dataset Construction

The experimental material was based on RadEvalX [4], a dataset designed to study the agreement between automatic metrics and radiologist assessment in chest X-ray report generation. RadEvalX contains 100 examples, each consisting of a reference report, a candidate report generated by the M2Tr model [6], and expert annotation of candidate-report errors with respect to the reference. The reference reports originate from the IU X-Ray collection [7]. The annotation covers eight error categories, including false and omitted findings, incorrect location or severity, errors involving comparison with prior studies, and errors involving uncertainty. Each error is additionally marked as clinically significant or clinically insignificant.

Since RadEvalX is originally in English, we constructed a Russian-language evaluation set for the present study. Both reference and candidate reports were translated into Russian using a large language model. To improve terminological consistency, an English–Russian terminology table for chest imaging was compiled from authoritative thoracic imaging terminology sources and used in the translation prompt [9,14]. Additionally, the study identifiers provided in RadEvalX were used to retrieve the corresponding chest X-ray images from IU X-Ray and align each translated report pair with its image. An example of a translated report is shown in Table 1.

The resulting dataset contained 100 aligned records consisting of: chest X-ray image, Russian reference report, Russian candidate report, total number of clinically significant errors, and total number of clinically insignificant errors.

Table 1. Example of an English radiology report and its Russian translation.

English (RadEvalX)
the heart size is within normal limits . trachea is midline . no pleural effusions or pneumothorax . cardiomediastinal contours are normal . there is focal consolidation in the posterior segment of the right lower lobe . no bony or soft tissue abnormalities .
Russian (Translation)
Сердце нормальных размеров. Трахея по средней линии. Плеврального выпота и пневмоторакса нет. Контуры сердца и отделов средостения обычные. Очаговая консолидация в заднем сегменте нижней доли правого лёгкого. Костные структуры и мягкие ткани без патологических изменений.

The count of clinically significant errors was used as the main target for fitting and evaluating the proposed metric, while the total error count was retained for supplementary analysis.

4 Adaptive Evaluation Framework

4.1 Framework Overview

The proposed framework is an end-to-end procedure for constructing and validating an evaluator under limited expert annotation. The resulting evaluator takes as input a candidate report, a reference report, and the corresponding chest X-ray image. It consists of three complementary components: an interpretable clinical component that identifies errors in the candidate report, a text-only component that compares the candidate and reference reports, and an image-text component that compares the candidate report with the radiograph. The final report-level score is obtained by combining the scores produced by these three components. The construction of the evaluator, namely the selection of the base components and the estimation of their fusion weights, is performed on an annotated training set whose records contain a chest X-ray image, a Russian reference report, a Russian candidate report, and an expert assessment of the candidate report. This selection-and-fitting procedure is validated without reusing the same examples for model selection and final testing. The design is motivated by the lack of Russian expert-annotated data: instead of training a full evaluator from scratch or manually fixing one metric, the framework adapts existing clinical, textual, and multimodal signals to the available supervision.

4.2 Candidate Components

The component pool is divided into three categories. Each possible configuration contains one reference-based Russian-language clinical component, one reference-based textual component, and one image-text alignment component.

The first component is a Russian clinical consistency score based on the GREEN evaluation framework [16]. In the Russian setting, GREEN was implemented using Qwen3.5-9B [18] without additional training on Russian medical reports. This component compares the reference and candidate reports and produces an interpretable evaluation summary.

The second group consists of reference-based textual components. Because many radiology-specific metrics are designed for English, Russian reports are translated into English before applying these methods. Three translation backbones were considered: TranslateGemma-12B-IT [8], HY-MT1.5-7B [24], and Qwen3.5-9B [18]. Since HY-MT1.5-7B and Qwen3.5-9B allow terminology usage, variants with and without the terminology table were evaluated. The textual candidate scores included RadGraph, RadCliQ, RaTEScore, along with embedding-based cosine similarity and BERTScore variants computed with radiology-specific encoders such as CXR-BERT-specialized and BioViL-T [2,3]. For metrics in which lower values indicate higher report quality, the reciprocal value was used.

The third group consists of image-text alignment components. These components were included to account for whether the candidate report is consistent with the corresponding chest X-ray image. One such signal was computed using BioViL-T image and text representations using cosine similarity.

4.3 Adaptive Score Fusion

For each possible configuration, the framework selects one Russian clinical component, one textual component, and one image-text component. The corresponding raw component values are written as f_{ru}, f_{text}, and f_{img}. Since different components have different scales, each value is standardized on the training subset of the current split:

$$z_{j,i}^{(c)} = \frac{f_{j,i}^{(c)} - \mu_j}{\sigma_j}, \qquad j \in \{ru, text, img\}. \tag{1}$$

Here, μ_j and σ_j are the mean and standard deviation of component j, respectively, estimated from that subset.

Let z_{ru}, z_{text}, and z_{img} denote the resulting standardized component values. The final quality score is defined as

$$S = w_{ru} z_{ru} + w_{text} z_{text} + w_{img} z_{img}, \tag{2}$$

where $w_{ru} > 0$, $w_{text} \geq 0$, $w_{img} > 0$, and $w_{ru} + w_{text} + w_{img} = 1$. Standardization parameters were always estimated on the corresponding training set and then applied to the validation or test set in order to prevent data leakage.

The target variable is defined as the number of clinically significant errors in the candidate report relative to the reference report. Since a higher value of S should indicate a better report whereas a higher target value indicates more errors, the optimization criterion was

$$J(c, w) = -\tau(S^{(c,w)}, y), \tag{3}$$

where τ is Kendall's rank correlation coefficient, c is the component configuration, w is the weight vector, and y is the expert error count. The weights were selected by grid search over the constrained two-dimensional simplex with step h:

$$\mathcal{W}_h = \{w \mid w_{ru} > 0,\ w_{text} \geq 0,\ w_{img} > 0, \\ w_{ru} + w_{text} + w_{img} = 1,\ \frac{w_j}{h} \in \{0, 1, 2, \ldots\}\}. \tag{4}$$

4.4 Configuration Selection under Limited Annotated Data

Because the same small dataset is used both to choose the evaluator and to estimate its quality, the framework uses nested cross-validation as a part of the selection procedure. In each outer cross-validation split, only the outer training set is available for choosing the component configuration and fusion weights. Within that training set, an inner cross-validation loop evaluates every possible configuration. For each inner fold, standardization parameters are estimated on the inner training subset, the fusion weights are selected from $\mathcal{W}_h$ on that subset using the objective in Eq. (3), and the resulting score is evaluated on the inner validation subset. The selected configuration is

$$c^* = \arg\max_{c \in \mathcal{C}} \overline{J}_{inner}(c), \tag{5}$$

where $\overline{J}_{inner}(c)$ is the mean inner-validation objective. After the configuration has been selected, standardization parameters and fusion weights are recomputed on the complete outer training split, and the resulting evaluator is applied once to the held-out outer test split. If the validation score is acceptable, the final evaluator can be obtained by training it on the whole dataset, or by picking any of the models from the outer cross-validation loop.

5 Experimental Evaluation

5.1 Implementation Details

The evaluation used all 100 report-image studies described in the dataset section. The resulting component pool contained one fixed Russian clinical component, 35 textual scoring components, and 5 image-text scoring components. Textual and image-text scores were matched by the translation model and terminology-table setting, i.e. the values computed from different English translations were not mixed; this allows one to reuse the same translation result for both evaluation branches during inference. This matching produced 35 possible configurations. At each training stage, the procedure selected one configuration of the form

(interpretable component, textual component, multimodal component)

and then optimized its fusion weights.

The nested procedure was conducted with $K_{\text{outer}} = 5$ outer folds and $K_{\text{inner}} = 5$ inner folds. The weight grid step was $h = 0.05$.

5.2 End-to-End Framework Evaluation

The nested cross-validation results are summarized in Table 2. The mean value of the objective function on the outer test folds was $J_{outer} = 0.5927$, with a standard deviation of 0.1491 and a 95% confidence interval of $[0.4621, 0.7234]$. Since $J = -\tau$, this corresponds to an average Kendall correlation of $\tau = -0.5927$ between the final quality score and the number of clinically significant errors. Thus, higher automatic scores were consistently associated with fewer expert-annotated errors.

Table 2. Nested cross-validation performance and learned fusion weights.

Quantity	Value
Outer-fold $J = -\tau$ values	0.5064, 0.4265, 0.7503, 0.5296, 0.7508
Mean J_{outer}	0.5927
Standard deviation	0.1491
95% confidence interval	$[0.4621, 0.7234]$
Mean weights	$w_{ru} = 0.11$, $w_{text} = 0.59$, $w_{img} = 0.30$
Weight ranges	$w_{ru} \in [0.05, 0.15]$, $w_{text} \in [0.45, 0.70]$, $w_{img} \in [0.15, 0.50]$

The learned weights show that the textual component was the dominant source of information on average. Nevertheless, the image-text component also received a substantial mean weight, confirming the value of incorporating visual consistency. The Russian clinical component received the smallest mean weight (still often above the minimum acceptable level), but it was retained in the selected solutions since it provides interpretability.

6 Discussion and Limitations

The results support the central design choice of the framework: in a low-resource Russian setting, evaluation should be treated as a selection-and-fusion procedure rather than as a single manually chosen metric. The nested cross-validation estimate and Fig. 1 show a stable negative correlation between the framework scores and the number of clinically significant expert-annotated errors.

The individual component analysis also clarifies the roles of the three signal families. Fig. 2, built on the whole dataset for each textual scoring component independently, shows that embedding-based semantic metrics (BioViL-T and CXR-BERT cosine similarity and BioViL-T-based BERTScore) were the most reliable individual English-language signals. This is plausible for radiology reports, where the same clinical finding may be expressed using different wording. The comparatively weaker correlation value of RadGraph F1 suggests that exact extraction and matching of entities and relations can become brittle after translation and under small-sample conditions. The nonzero contribution of the

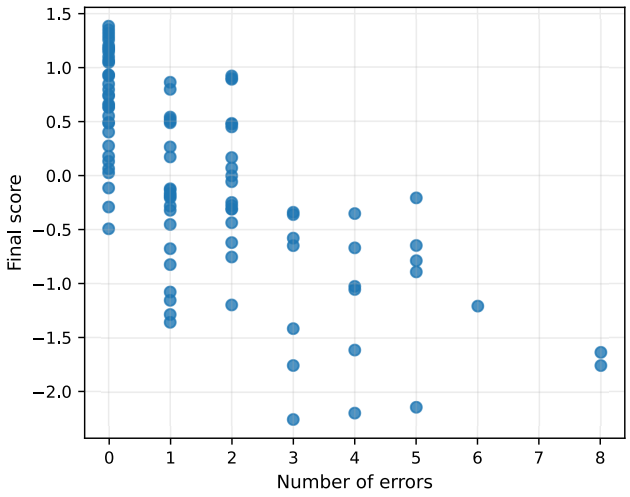

Fig. 1. Final framework score versus the number of clinically significant errors on the union of the outer test folds. Higher framework scores correspond to fewer expert-annotated errors.

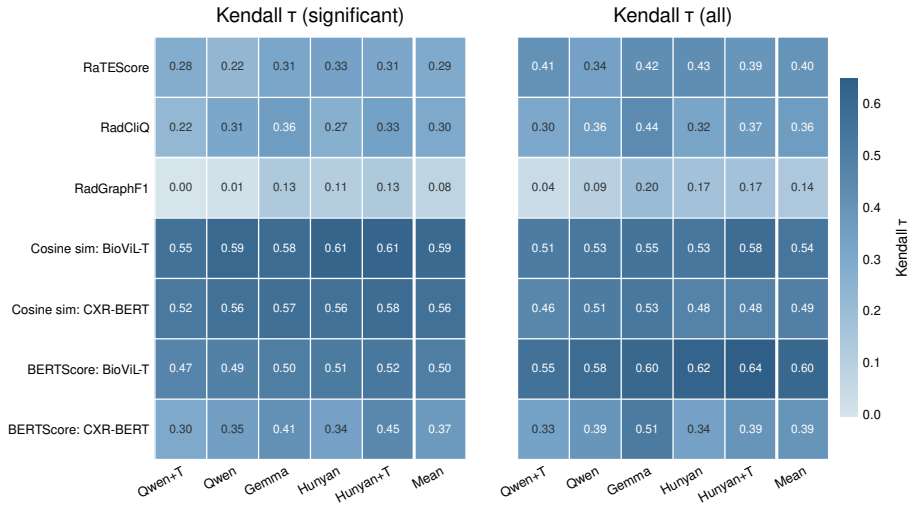

Fig. 2. Kendall correlation between textual component scores and inverse expert error counts. The left panel uses clinically significant errors, while the right panel uses all errors. Columns correspond to translation and terminology configurations, the suffix "+T" denotes the use of terminology in the prompt. The final column reports the mean of the first 5 columns.

image-text component indicates that reference-based comparison alone is incomplete: a candidate report should also remain aligned with its X-ray image.

The Russian clinical component received the smallest average weight, but this should not be interpreted as evidence that clinical reasoning is unimportant. In the present implementation, the component uses a general open language model without Russian radiology fine-tuning. Its main practical value is therefore twofold: it contributes a scalar signal to the ensemble and can provide interpretable descriptions of report discrepancies, which are useful for diagnosing model failures.

Several limitations should be considered. First, the evaluation set contains only 100 studies, so the confidence interval is wide and high variance is expected. Second, the Russian reports were produced by automatic translation, and the translated dataset has not yet been independently validated by physicians. Translation artifacts may affect performance of all 3 components. Third, most specialized components were originally developed for English radiology reports, so their use in Russian depends on the quality of translation and terminology preservation. Fourth, the framework was evaluated on RadEvalX-derived examples with candidate reports generated by a specific model; additional validation is required for other Russian corpora, reporting styles, and generation systems. Fifth, the Russian clinical component relies on a general-domain LLM without Russian radiology fine-tuning, which may underestimate its true potential. Future work should fine-tune on Russian chest X-ray reports. Finally, the framework is intended for automatic evaluation and model comparison, not for clinical decision making.

7 Conclusion

We presented an adaptive framework for evaluating Russian-language chest X-ray reports under limited annotated data. The framework constructs an evaluator by combining a Russian clinical fidelity signal, translated reference-based textual metrics, and image-report alignment scores, then selecting the component configuration and fusion weights through a nested cross-validation procedure. On 100 aligned RadEvalX-based studies, the framework achieved an average Kendall correlation of -0.5927 between the final quality score and clinically significant error counts.

The study demonstrates that clinically oriented evaluation for a non-English radiology setting can be approached by adapting and fusing pretrained text and vision-language models. Future work should focus on physician validation of the translated Russian dataset, evaluation on larger Russian radiology benchmarks, and adaptation to local reporting conventions.

Acknowledgements The study was conducted under the state assignment of Lomonosov Moscow State University.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (2005)
2. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15016–15027 (2023)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 1–21. Springer Nature Switzerland, Cham (2022)
4. Calamida, A.R., Nooralahzadeh, F., Rohanian, M., Nishio, M., Fujimoto, K., Krauthammer, M.: Radiology Report Generation Models Evaluation Dataset For Chest X-rays (RadEvalX). PhysioNet (2024), `https://doi.org/10.13026/tp88-q278`, version 1.0.0. Accessed: 2026-05-31
5. Cho, G., Jang, S., Ko, H., Baek, I., Park, C.M.: CREPE: Rapid chest X-ray report evaluation by predicting multi-category error counts. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 21738–21755. Association for Computational Linguistics, Suzhou, China (2025)
6. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10578–10587 (2020)
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
8. Finkelstein, M., Caswell, I., Domhan, T., Peter, J.T., Juraska, J., Riley, P., Deutsch, D., Kovacs, G., Dilanni, C., Cherry, C., Briakou, E., Nielsen, E., Luo, J., Black, K., Mullins, R., Agrawal, S., Xu, W., Kats, E., Jaskiewicz, S., Freitag, M., Vilar, D.: TranslateGemma technical report (2026), `https://arxiv.org/abs/2601.09012`, accessed: 2026-05-31
9. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Müller, N.L., Remy, J.: Fleischner society: Glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)
10. Huang, A., Banerjee, O., Wu, K., Reis, E.P., Rajpurkar, P.: FineRadScore: A radiology report line-by-line evaluation technique generating corrections with severity scores (2024), `https://arxiv.org/abs/2405.20613`, accessed: 2026-05-31
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 590–597 (2019)

12. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P.: RadGraph: Extracting clinical entities and relations from radiology reports (2021), https://arxiv.org/abs/2106.14463, accessed: 2026-05-31
13. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
14. Nikolaev, A.E., Suchilova, M.M., Korkunova, O.A., Blokhin, I.A., Gonchar, A.P., Reshetnikov, R.V., Morozov, S.P.: Terminologiya opisaniya organov grudnoi kletki — rentgenografiia i komp'iuternaia tomografiia: metodicheskie rekomendatsii [Terminology for chest organ description: radiography and computed tomography: guidelines] (in Russian). GBUZ "NPKTs DiT DZM", Moskva (2022)
15. Nsengiyumva, N.P., Hussain, H., Oxlade, O., Majidulla, A., Nazish, A., Khan, A.J., Menzies, D., Ahmad Khan, F., Schwartzman, K.: Triage of persons with tuberculosis symptoms using artificial intelligence–based chest radiograph interpretation: A cost-effectiveness analysis. Open Forum Infectious Diseases **8**(12), ofab567 (2021)
16. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., 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 (2024)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002)
18. Qwen Team: Qwen3.5: Towards native multimodal agents (2026), https://qwen.ai/blog?id=qwen3.5, accessed: 2026-05-31
19. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519. Association for Computational Linguistics, Online (2020)
20. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., Langlotz, C.P., Venugopal, V.K., Rajpurkar, P.: Evaluating progress in automatic chest X-ray radiology report generation. Patterns **4**(9) (2023)
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert (2020), https://arxiv.org/abs/1904.09675, accessed: 2026-05-31
22. Zhang, Z., Lee, K., Jing, P., Deng, W., Zhou, H., Jin, Z., Huang, J., Gao, Z., Marshall, D.C., Fang, Y., et al.: GEMA-Score: Granular explainable multi-agent scoring framework for radiology report evaluation. Proceedings of the AAAI Conference on Artificial Intelligence **40**(15), 13025–13033 (2026)
23. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: RaTEScore: A metric for radiology report generation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 15004–15019. Association for Computational Linguistics, Miami, Florida, USA (2024)

24. Zheng, M., Li, Z., Chen, T., Song, M., Wang, D.: HY-MT1.5 technical report (2025), https://arxiv.org/abs/2512.24092, accessed: 2026-05-31