Enhancing Human-Computer Interaction in Chest X-ray Analysis using Vision and Language Model with Eye Gaze Patterns

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Yue Gao, and Honghan Wu

University College London
yunsoo.kim.23@ucl.ac.uk

Abstract. Recent advancements in Computer Assisted Diagnosis have shown promising performance in medical imaging tasks, particularly in chest X-ray analysis. However, the interaction between these models and radiologists has been primarily limited to input images. This work proposes a novel approach to enhance human-computer interaction in chest X-ray analysis using Vision-Language Models (VLMs) enhanced with radiologists' attention by incorporating eye gaze data alongside textual prompts. Our approach leverages heatmaps generated from eye gaze data, overlaying them onto medical images to highlight areas of intense radiologist's focus during chest X-ray evaluation. We evaluate this methodology in tasks such as visual question answering, chest X-ray report automation, error detection, and differential diagnosis. Our results demonstrate the inclusion of eye gaze information significantly enhances the accuracy of chest X-ray analysis. Also, the impact of eye gaze on fine-tuning was confirmed as it outperformed other medical VLMs in all tasks except visual question answering. This work marks the potential of leveraging both the VLM's capabilities and the radiologist's domain knowledge to improve the capabilities of AI models in medical imaging, paving a novel way for Computer Assisted Diagnosis with a human-centred AI. The code for processing data and evaluation can be found at https://github.com/knowlab/CXR_VLM_EyeGaze.

Keywords: Vision Language Model · Eye Gaze · Chest X-ray

1 Introduction

Recent AI breakthroughs have facilitated the development of advanced diagnostic tools such as Computer-Aided Diagnosis (CAD). These systems leverage machine learning and deep learning models to tackle medical challenges [8, 21, 23]. While CAD demonstrated promising results in improving diagnostic accuracy, concerns remain about its reliability and effectiveness as a standalone tool in clinical settings.

One solution is leveraging human-computer interaction. Studies show integrating human expertise into CAD enhanced accuracy and reliability. This

approach outperformed both radiologists and AI models making the decision alone in diagnostic accuracy [3, 20]. However, the current AI models used with human-computer interaction for medical image analysis are dominantly limited to analyzing images only, thereby restricting the usage of these models in clinical settings.

A recent breakthrough in Vision-Language Models (VLMs) with Large Language Models (LLMs) extends CAD applications with human-computer interaction to complex multimodal data. This significant advancement in the interpretation of medical images and reports enables the analysis of diagnostic images, such as chest X-rays (CXRs), through textural prompts which can include indications, reports, and any other text inputs that we want the model to leverage. These models have shown strong performance in unseen tasks, and this versatility and robustness highlight the potential of these models as de-facto models for CAD [12, 19, 24].

For CXRs, VLMs have demonstrated utility in several tasks, including the automatic generation of radiology findings from images, visual question answering of these images, and correction of radiology reports based on CXR images [28–30]. Through these capabilities, VLMs not only streamline the diagnostic workflow but also offer valuable insights that can aid radiologists in making informed decisions. Although these models are designed to serve as interactive assistants, the human-computer interaction is limited to input CXR images and text prompts.

To further enhance human-computer interaction in CXR analysis using VLMs, we propose a novel method of integrating eye gaze data into the VLM framework for CXRs. This approach utilizes heatmaps generated from eye gaze patterns recorded during the interpretation of images. By incorporating these heatmaps, we introduce an additional layer of insight to the VLM, representing an advancement in human-centred AI in medical image analysis. We also further fine-tuned the models with eye gaze data. In this work, we tested our models' effectiveness in four clinical tasks: Report Automation, Error Detection, Differential Diagnosis, and Visual Question and Answering.

Our contributions in this paper are as follows:

1. **Enhancing Human-Computer Interaction for VLMs with Eye Gaze**: This work used heatmaps with VLMs for clinical applications for the first time, which highlight the precise focal points and duration of a radiologist's attention when analysing a CXR.
2. **Comprehensive evaluation on 4 real-world clinical applications**: We test the effectiveness of our approach in a comprehensive list of models (10 models including our own fine-tuned model) on 4 clinical applications: Report Automation, Error Detection, Differential Diagnosis and Visual Question and Answering.

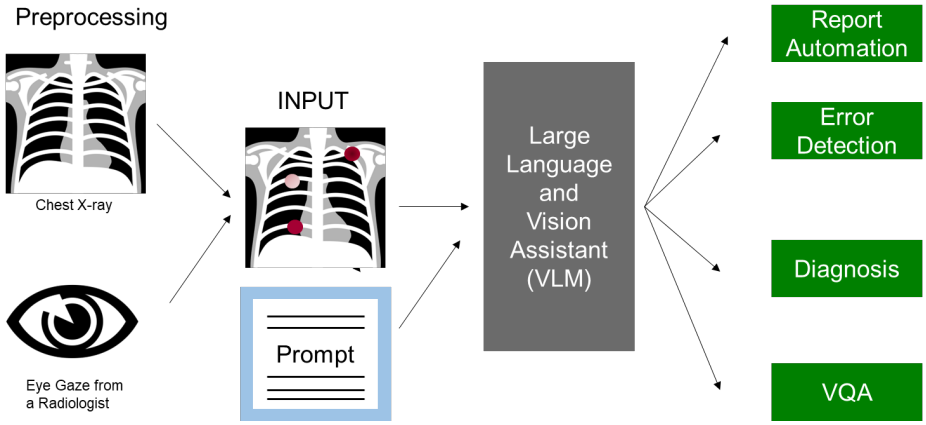

Fig. 1. Overview of Enhancing Human-Computer Interaction in Chest X-ray Analysis using Vision and Language Model with Eye Gaze Patterns.

2 Related Work

Several studies have shown promising results by incorporating radiologists' eye-tracking data into AI models, leading to improved diagnostic accuracy [18, 9, 26, 31]. For example, the EG-ViT model leverages radiologists' eye gaze data to guide its attention towards potentially pathological regions [18]. Another model Mammo-Net explores enhancing mammogram classification by incorporating eye gaze data and multi-view information to improve mammogram classification, which addresses interpretability and data annotation limitations by using radiologists' eye gaze data [9]. However, these models and other works typically focus on analyzing only images, which constrains the usage of the model to single-modality applications [26, 31].

Textual data, such as clinical notes, are equally critical in the clinical setting, containing detailed and comprehensive information. Thus, we believe embedding both images and reports in Visual Language Models (VLMs) can achieve a more holistic understanding and improve diagnostic accuracy.

3 Tasks

Figure 1 summarizes an overview of our CXR analysis framework utilizing eye gaze patterns in the following 4 tasks in CAD. This figure shows how the input CXR image is processed to incorporate eye gaze patterns. Paired with textual prompts for the downstream task, the eye gaze heatmap-enhanced CXR images provide extra human intelligence to the VLM.

3.1 Report Automation (GEN and SUM)

Radiology report automation typically involves generating and summarizing radiology reports through VLMs. A standard radiology report includes a "Findings" section, which describes the observations made from the images, and an "Impressions" section, which offers a summary for diagnostic purposes. The task of report generation, called hereafter as **GEN**, aims to produce the "Findings" section based on the images. In contrast, report summarization, called hereafter as **SUM**, seeks to construct the "Impressions" section by leveraging both the "Findings" and the images.

3.2 Error Detection (ERR)

In light of errors found in radiologists' reports, ensuring diagnostic accuracy is imperative [2]. It is shown by other work leveraging VLMs as supportive tools can help assist radiologists in identifying errors that appeared in reports [28]. For such error detection tasks, we created the evaluation data by introducing synthetic errors into the original report to test VLM's ability to classify their presence ("Y") or absence ("N"). Specifically, with the help of radiologists, we identified the top 33 most important phenotypes that must be reported in reports if present in corresponding CXR. So, when more than one of such phenotypes exists, we randomly choose a phenotype and introduce an error based on it, and some of the reports were left unchanged [28].

3.3 Differential Diagnosis (DDx)

In this task, the model generates potential diagnoses from chest X-ray images. Given that diagnoses are commonly classified using the International Classification of Diseases (ICD) codes, we refine the model's raw text output to correspond with these classifications. For disease entity recognition within the text, we employ a DeBERTa-V3-large model, fine-tuned on the BC5CDR dataset [5, 25, 27]. To ensure accurate disease entity alignment with ICD codes, we utilize embeddings from a SapBERT model [14].

3.4 Visual Question Answering (VQA)

To evaluate the effectiveness of current state-of-the-art VLMs on the reasoning and understanding of clinical knowledge, we used the MIMIC-CXR-VQA dataset for such evaluation [1]. It is an image-based Electronic Health Records (EHR) question-answering dataset that is designed to facilitate joint reasoning across imaging and table modalities in EHR question-answering (QA) systems.

4 Methods

4.1 Datasets

In this study, we leverage the posterior to anterior (PA) view images from the MIMIC-Eye dataset, a compilation of MIMIC-IV, MIMIC-CXR, REFLACX,

and EyeGaze datasets [6]. This dataset is comprised of 3,689 chest X-ray images for intensive care unit patients. REFLACX and EyeGaze provide eye-tracking data for radiologists, which is an essential component of our work. While the EyeGaze dataset provides a list of diagnoses, the REFLACX dataset does not provide a list of diagnoses, making the REFLACX dataset unfit for the DDx task.

For the evaluation dataset, We further processed the EyeGaze dataset as it can be used for all the tasks in our study. We use the overlap of the EyeGaze dataset with MIMIC-CXR-VQA and reports with synthetic errors for the **ERR** task, 574 images. The rest of the EyeGaze dataset and REFLACX dataset were combined to form a train dataset, 1,169 images. We ensured there was no overlap between the evaluation and train dataset. A detailed breakdown of the train and evaluation datasets is provided in Table 1.

For all the datasets, raw eye gaze information at 1000 Hz is used to create a heat map on the CXR image. We drew a red dot at the x and y coordinate of the gaze with the darkness of the dot representing the number of gazes. In other words, darker dots show more time the radiologist spent time on the spot of the CXR image.

Table 1. Train and Evaluation Dataset Description.

Statistics	**Train**	**Evaluation**
Number of CXR Images	1,169	574
CXR Images with DDx cases	420	574
Number of Visual Questions	1,542	574
Number of Reports	940	574
Reports Error Rates	73.09%	73.17%
Total Number of Instruction Data	4,782	N/A

4.2 List of Models

Table 2 shows the models we use for comparison. We include LLaVA variants from open domain (LLaVA-v0 [17], LLaVA-v1.5 [15] and LLaVA-v1.6 [16]) and medical domain (LLaVA-Med [11] and CXR-LLaVA [10]). All models except LLaVA-v1.5 and LLaVA-1.6 are based on 7B backbone LLM and LLaVA-v1.5 and LLaVA-1.6 with two model sizes: 7B and 13B. LLaVA-1.6 comes with an additional backbone LLM, Mistral-7B.

4.3 Model Train

Fine-tuning code is from LLaVA's official GitHub repository. We kept the LLaVA's default hyperparameter configurations with two adjustments: the train batch size and the number of epochs. We reduced the train batch size to 8 and limited the number of epochs to 1.

Table 2. Model Description.

Model	Vision Encoder	Resolution (pixel)	Backbone LLM	Connector	Train Data Size
LLaVA-v0	CLIP ViT-L/14	224	Vicuna-7B-v0	Projection	753K
LLaVA-Med	CLIP ViT-L/14	224	Vicuna-7B-v0	Projection	560K
LLaVA-v1.5-7B	CLIP ViT-L/14	336	Vicuna-7B-v1.5	MLP	1223K
LLaVA-v1.5-13B	CLIP ViT-L/14	336	Vicuna-13B-v1.5	MLP	1223K
LLaVA-v1.6-7B	CLIP ViT-L/14	672	Vicuna-7B-v1.5	MLP	1318K
LLaVA-v1.6-13B	CLIP ViT-L/14	672	Vicuna-13B-v1.5	MLP	1318K
LLaVA-v1.6M	CLIP ViT-L/14	672	Mistral-7B	MLP	1318K
CXR-LLaVA	CXR-specific CLIP ViT-L/16	512	LLaMA-7B	Projection	527K

Training our model with 3 A5000 GPUs presented memory limitations. To overcome this challenge, we implemented several techniques that significantly reduced memory consumption. These techniques included low-rank adaptation (LoRA), the DeepSpeed zero-redundancy optimizer (ZeRO3), and flash attention [7, 22, 4].

We trained two versions of one of the open domain models, LLaVA-v1.5-7B: one with the raw CXR images and another one with the eye gaze pattern CXR images. We denote this model with 'FT' and 'FT+G' respectively.

4.4 Evaluation

To ensure efficient and consistent evaluations throughout our experiments, we adopted a zero-shot approach for all tasks and used a batch of 1 with a temperature parameter of 0. We opted temperature to be 0 to minimize the randomness of the model's generated text. For each task, we set specific limits on the maximum length of the model's response: **GEN** - 320, **SUM** - 128, **ERR** - 64, **DDx** - 192, **VQA** - 64. These limits were chosen based on the expected length of a typical response in each task. This optimization helps the model perform better and ensures our evaluation results are reliable and consistent across different tasks. To test the effectiveness of the eye gaze pattern, we also evaluated the model with the raw CXR images ('No Gaze') and with the CXR images with the eye gaze pattern ('Gaze')

We use different evaluation metrics for different tasks. For report automation including **GEN** and **SUM**, we use the ROUGE score [13]. It measures the overlap of n-grams, word sequences, and sometimes word pairs between the generated summary and the reference summaries. Among various versions of the ROUGE score, we chose to use ROUGE-L because it focuses on the longest common subsequence. We used HuggingFace's `evaluate` package to calculate the score.

For the **ERR** task, we utilize the accuracy score to evaluate this binary classification performance. To extract a valid answer, we apply regular expression and `thefuzz` package for string matching.

For **DDx**, the calculation of the F1 score is adapted to accommodate the specificity of the task. Diagnosis predictions are extracted from the model's responses, focusing on the accuracy at the ICD code level through disease entity recognition and alignment. Precision is computed by dividing the count of correct predictions by the total number of predictions made, while recall is determined by the ratio of correct predictions to the total number of relevant diseases for the patient. These values of recall and precision are then employed to compute the F1 score. For the **VQA** task, we employ the accuracy score as the metric to assess performance.

5 Results and Discussion

Table 3. Evaluation Results. **GEN**: Report Genetaion. **SUM**: Report Summarization. **ERR**: Error Detection. **DDx**: Differential Diagnosis. **VQA**: Visual Question Answering. The evaluation metrics used in each task are noted in parentheses. No G and G stands for 'No Gaze' and 'Gaze.' We bold the scores in 'Gaze' if they are higher than the corresponding 'No Gaze' scores.

Model	**Size**	**GEN** (R-L)		**SUM** (R-L)		**ERR** (Acc)		**DDx** (F1)		**VQA** (Acc)	
		No G	**G**	**No G**	**G**	**No G**	**G**	**No G**	**G**	**No G**	**G**
LLaVA-v0	7B	9.86	8.72	9.12	8.89	28.75	**71.78**	2.70	**4.59**	43.03	40.07
LLaVA-Med	7B	11.39	**12.41**	9.99	9.64	70.21	**73.00**	2.59	**11.10**	46.69	43.03
LLaVA-v1.5	7B	15.10	13.56	10.68	9.90	49.13	37.80	2.62	**6.15**	47.74	44.77
CXR-LLaVA	7B	24.88	24.60	39.25	**41.43**	48.26	**48.61**	12.35	**13.31**	56.79	**59.06**
LLaVA-v1.6	7B	11.27	10.04	10.09	10.00	56.27	49.13	7.41	**10.48**	44.77	**45.30**
LLaVA-v1.6M	7B	12.52	11.53	10.23	9.82	58.89	44.95	4.33	**7.05**	56.45	**58.54**
LLaVA-v1.5	13B	14.97	11.44	11.19	10.46	35.54	28.92	4.25	**5.64**	45.30	**46.34**
LLaVA-v1.6	13B	11.86	11.41	10.33	10.32	67.42	59.23	5.25	**6.53**	42.16	42.16
LLaVA-v1.5FT	7B	29.52	29.30	53.93	53.89	73.87	73.87	23.69	23.69	61.67	61.67
LLaVA-v1.5FT+G	7B	29.32	**29.76**	53.19	**53.25**	74.39	74.39	18.16	18.16	52.79	**57.49**

Analysing the inference performance of VLMs with eye gaze patterns with those without highlighted intriguing insights in these models as well as the incorporation of this extra human intelligence. The comparison evaluation results are detailed in Table 3.

5.1 Eye gaze enhances DDx most evidently

All the baseline models perform better with eye gaze patterns in DDx. The largest increase from 'No Gaze' (2.59%) to 'Gaze' (11.10%) was seen by the **LLaVA-Med model**. This increase in performance is not seen in all the models in other tasks. Still, we observe the improvement in at least one baseline model in all the tasks. In the error detection task, we see the largest increase in **LLaVA-v0**, a 43.03% increase from 'No Gaze' (28.75%) to 'Gaze' (71.78%).

LLaVA-Med and **CXR-LLaVA** performance also increased with eye gaze patterns. Apart from these models, **LLaVA-v1.6**, **LLaVA-v1.6M**, and **LLaVA-v1.5-13B** models saw an increase in performance for the **VQA** task. These findings highlight the effectiveness of incorporating this additional human-computer interaction to enhance model performance.

5.2 Medical models perform better with eye gaze patterns

The evaluation result also highlighted significant performance enhancements in models fine-tuned within the medical domain data, particularly **LLaVA-Med** and **CXR-LLaVA**. **LLaVA-Med**, which was fine-tuned with PMC figure and legends, show an increase in performance with eye gaze pattern for **GEN**, **ERR**, and **DDx**. **CXR-LLaVA** model, which was fine-tuned with MIMIC reports, saw an increase in performance with the 'Gaze' image in all tasks except **GEN**. The result suggests that domain-specific fine-tuning can significantly enhance VLM performance in clinical applications and may play a role in this improvement of performance with eye gaze patterns.

5.3 Larger models do not perform well

Interestingly, our evaluation revealed that larger models did not consistently perform better in some tasks. The LLaVA-v1.5-13B model performed worse than its 7B model in **GEN**, **ERR**, and **VQA**. Also, this trend is observed in LLaVA-v1.6 models: 7B model outperforming in **DDx** and **VQA**. This result suggests that adding more parameters does not directly translate to performance improvements.

5.4 Impact of eye gaze patterns in fine-tuning

We selected the LLaVA-v1.5-7B model because it was the best-performing model apart from the LLaVA-v1.6 models, for which the training code was not publicly available. Both versions of LLaVA-v1.5-7B fine-tuned models exhibited superior performance over all the other models, including the medical models, except for the **VQA** task. Also, the eye gaze pattern proved to be promising in enhancing the model's performance for **GEN** and **ERR** tasks which the baseline model actually showed a decrease in performance.

These findings pave the way for further exploration of eye gaze data integration in VLM-based clinical applications. By leveraging radiologists' eye gaze data, we showed promise to enhance VLMs for more accurate and insightful clinical decision support systems in medical imaging analysis.

While our findings indicate that eye gaze data enhances model performance for specific tasks such as DDx, the improvement is not uniformly observed across other tasks (GEN and SUM). Additionally, some results remained unchanged regardless of the inclusion of gaze data. These inconsistencies underscore the complexity of integrating human cognitive data into automated systems. Our work,

therefore, highlights both the potential benefits and the challenges of leveraging eye gaze patterns, offering a critical perspective that will inform and inspire future research on integrating eye gaze data in LLM-based VLMs.

Future research could also delve deeper into the interplay between eye gaze patterns and model architectures to unlock the full potential of these collaborative AI systems in healthcare. Also, in the future, this work can be extended to other types of medical datasets such as CT and MRI as well as other tasks in CAD such as segmentation.

6 Conclusion

In conclusion, our study has demonstrated the potential of integrating radiologists' eye-tracking data into VLMs to enhance the accuracy of analyzing chest X-rays. This is particularly evident in the area of differential diagnosis, where all the baseline models we tested achieved a remarkable improvement with eye gaze patterns. Also, the positive impact of eye gaze data on fine-tuned models suggests its potential to enhance performance even in tasks where the baseline model struggles. This enhanced human-computer interaction approach promises to improve decision-making in clinical practices, suggesting a pivotal step toward a more synergistic human-AI collaboration.

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

References

1. Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E., Kim, T., et al.: Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. Advances in Neural Information Processing Systems **36** (2024)
2. Brady, A.P.: Error and discrepancy in radiology: inevitable or avoidable? Insights into imaging **8**, 171–182 (2017)
3. Calisto, F.M., Santiago, C., Nunes, N., Nascimento, J.C.: Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. Artificial Intelligence in Medicine **127**, 102285 (2022)
4. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)
5. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
6. Hsieh, C., Ouyang, C., Nascimento, J.C., Pereira, J., Jorge, J., Moreira, C.: Mimic-eye: Integrating mimic datasets with reflacx and eye gaze for multimodal deep learning applications (2023)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

8. Hwang, E.J., Lee, J.H., Kim, J.H., Lim, W.H., Goo, J.M., Park, C.M.: Deep learning computer-aided detection system for pneumonia in febrile neutropenia patients: a diagnostic cohort study. BMC Pulmonary Medicine **21**(1), 406 (2021). https://doi.org/10.1186/s12890-021-01768-0, https://doi.org/10.1186/s12890-021-01768-0
9. Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., Shen, D.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–78. Springer (2023)
10. Lee, S., Youn, J., Kim, M., Yoon, S.H.: Cxr-llava: Multimodal large language model for interpreting chest x-ray images. arXiv preprint arXiv:2310.18341 (2023)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
12. Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., Zhou, L.: A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. medRxiv pp. 2023–11 (2023)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. arXiv preprint arXiv:2010.11784 (2020)
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
16. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
18. Ma, C., Zhao, L., Chen, Y., Wang, S., Guo, L., Zhang, T., Shen, D., Jiang, X., Liu, T.: Eye-gaze-guided vision transformer for rectifying shortcut learning. IEEE Transactions on Medical Imaging (2023)
19. OpenAI: Gpt-4 (2023), https://www.openai.com/gpt-4
20. Patel, B.N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarappallil, J., Mariano, A.J., Riley, G., Seekins, J., Shen, L., Zucker, E., Lungren, M.P.: Human–machine partnership with artificial intelligence for chest radiograph diagnosis. npj Digital Medicine **2**(1), 111 (2019). https://doi.org/10.1038/s41746-019-0189-7, https://doi.org/10.1038/s41746-019-0189-7
21. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. BioMedical Engineering OnLine **17**(1), 113 (2018). https://doi.org/10.1186/s12938-018-0544-y, https://doi.org/10.1186/s12938-018-0544-y
22. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3505–3506 (2020)
23. Shaheed, K., Szczuko, P., Abbas, Q., Hussain, A., Albathan, M.: Computer-aided diagnosis of covid-19 from chest x-ray images using hybrid-features and random forest classifier. Healthcare **11**(6) (2023). https://doi.org/10.3390/healthcare11060837, https://www.mdpi.com/2227-9032/11/6/837

24. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. arXiv preprint arXiv:2307.14334 (2023)
25. Ushio, A., Camacho-Collados, J.: T-ner: an all-round python library for transformer-based named entity recognition. arXiv preprint arXiv:2209.12616 (2022)
26. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
27. Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., Lu, Z.: Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. Database **2016** (2016)
28. 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 preprint arXiv:2312.13103 (2023)
29. Wu, J., Kim, Y., Wu, H.: Hallucination benchmark in medical visual question answering. arXiv preprint arXiv:2401.05827 (2024)
30. Yildirim, N., Richardson, H., Wetscherek, M.T., Bajwa, J., Jacob, J., Pinnock, M.A., Harris, S., de Castro, D.C., Bannur, S., Hyland, S.L., et al.: Multimodal healthcare ai: Identifying and designing clinically relevant vision-language applications for radiology. arXiv preprint arXiv:2402.14252 (2024)
31. Zhao, Z., Wang, S., Wang, Q., Shen, D.: Mining gaze for contrastive learning toward computer-assisted diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7543–7551 (2024)