

Collaboration between clinicians and vision–language models in radiology report generation

Received: 8 February 2024

Accepted: 16 September 2024

Published online: 7 November 2024

 Check for updates

Ryutaro Tanno^{1,6}✉, David G. T. Barrett^{1,6}✉, Andrew Sellergren², Sumedh Ghaisas¹, Sumanth Dathathri¹, Abigail See¹, Johannes Welbl¹, Charles Lau², Tao Tu¹, Shekoofeh Azizi¹, Karan Singhal^{2,4}, Mike Schaekermann², Rhys May¹, Roy Lee², SiWai Man², Sara Mahdavi¹, Zahra Ahmed¹, Yossi Matias², Joelle Barral¹, S. M. Ali Eslami¹, Danielle Belgrave^{1,5}, Yun Liu², Sreenivasa Raju Kalidindi³, Shravya Shetty², Vivek Natarajan², Pushmeet Kohli¹, Po-Sen Huang¹, Alan Karthikesalingam²✉ & Ira Ktena¹✉

Automated radiology report generation has the potential to improve patient care and reduce the workload of radiologists. However, the path toward real-world adoption has been stymied by the challenge of evaluating the clinical quality of artificial intelligence (AI)-generated reports. We build a state-of-the-art report generation system for chest radiographs, called Flamingo-CXR, and perform an expert evaluation of AI-generated reports by engaging a panel of board-certified radiologists. We observe a wide distribution of preferences across the panel and across clinical settings, with 56.1% of Flamingo-CXR intensive care reports evaluated to be preferable or equivalent to clinician reports, by half or more of the panel, rising to 77.7% for in/outpatient X-rays overall and to 94% for the subset of cases with no pertinent abnormal findings. Errors were observed in human-written reports and Flamingo-CXR reports, with 24.8% of in/outpatient cases containing clinically significant errors in both report types, 22.8% in Flamingo-CXR reports only and 14.0% in human reports only. For reports that contain errors we develop an assistive setting, a demonstration of clinician–AI collaboration for radiology report composition, indicating new possibilities for potential clinical utility.

Radiology plays an integral and increasingly important role in modern medicine, by informing diagnosis, treatment and management of patients through medical imaging. However, the current global shortage of radiologists restricts access to expert care and causes heavy workloads for radiologists, resulting in undesirable delays and errors in clinical decisions^{1,2}. In the past decade, we have witnessed the remarkable promise of AI algorithms as assistive technology for improving the access, efficiency and quality of radiological care, with

more than 200 US Food and Drug Administration approved commercial products developed by companies based in more than 20 countries³ and approximately one in every three radiologists in the United States already benefiting from AI as part of their clinical workflow⁴.

The vast majority of these approved AI applications, however, focus only on the classification and quantification of very specific pathologies⁵. In practice, clinical radiology is much more than an accumulation of such narrow interpretive tasks, because findings must be

A full list of affiliations appears at the end of the paper. ✉e-mail: rtanno@google.com; barrettdavid@google.com; alankarthi@google.com; iraktena@google.com

communicated with appropriate nuance, synthesized in a broader clinical context and combined with overall impressions and recommendations that are useful for patient care. Radiologist experts use natural language to communicate this synthesis of the imaging findings alongside their overall impression and recommendations in the form of written reports. The recent progress in AI for modeling vision and language data simultaneously^{6–9}, coupled with the growing availability of digitized multimodal radiology data, has enabled the possibility of developing an automatic report generation system that is capable of producing a complete free-text description of the medical image^{10–14}. Framing report generation as the north star for a useful radiology AI system is more closely aligned to current radiologist practice and patient care, and allows for a more fine-grained and diverse description of the relevant findings that can be tailored to the needs of a given clinical scenario, including aspects such as location, size and severity, ambiguity, relation to clinical context of specific pathologies or their impact on onward care and more¹⁵.

Despite the increasing number of publications on AI-based report generation and its potential in improving the radiology workflow, automated report generation has not yet been widely adopted in real practice⁵. Several unmet needs represent key barriers to automated reporting achieving real-world impact. One notable obstacle is the difficulty of meaningfully evaluating the clinical quality of generated reports. The high degree of freedom in free-form reports introduces a wide range of possible errors to measure and classify. Exacerbating this, the desirable contents of a report differ between clinical settings (for example, an emergency setting versus a medical check-up), geographic regions¹⁶ and preferred approaches to standardization¹⁷. Previous works have approached this challenge by proposing automated metrics for evaluating the clinical quality of generated reports^{18–21} but many limitations remain. First, there has been a paucity of comprehensive evaluation of automated reports against reports produced by human experts (certified radiologists), which are known themselves to have variable style and quality. Despite impressive progress in automated metrics for report quality, only one study²² has directly assessed whether AI-generated reports were considered preferable to those by human experts, whereas others²³ have evaluated their utility in practice in a specific clinical setting only. Furthermore, the reasons given for preference choices have not been explored sufficiently. Second, previous work has only evaluated AI-generated reports as stand-alone artifacts, meaning the utility of these systems as assistive tools remains unknown. Evaluation in clinician–AI collaboration scenarios is arguably more realistic, given that most AI tools approved for clinical decision-making have been developed for an assistive rather than autonomous role in care delivery^{24,25}.

In addition to the above evaluation challenges, there remains considerable headroom for improvement in the clinical accuracy of existing AI report generation models²¹. Recent breakthroughs in multimodal foundation models^{9,21} have demonstrated that AI systems trained on a vast quantity of unlabeled data can be adapted and achieve state-of-the-art accuracy in a wide range of downstream specialized tasks, including biomedical problems²⁶. However, most existing report generation models^{10–13} are built from scratch, neglecting the likely useful transfer of knowledge from such pretrained models. By leveraging advances accrued through large-scale pretraining of vision–language models and tailoring them to a specific medical task, there is an opportunity to build an even more powerful report generation system.

In this work, we directly address these key unmet needs for AI report generation. We present Flamingo-CXR, a system for AI report generation predicated on a recent vision–language foundation model that achieves state-of-art performance in multiple automated metrics⁸. We evaluate Flamingo-CXR on historic, deidentified datasets across a diversity of clinical and geographic settings—both intensive care in the United States and in/outpatient care delivery in India—and move

beyond automated metrics to a detailed human evaluation of the reports generated with a pool of 27 radiologists, including a direct comparison of clinicians' preferences for AI reports versus human reports. Furthermore, we evaluate the system in an autonomous as well as assistive context. Figure 1 shows an overview of the proposed evaluation framework.

Our contributions richly characterize the wide spectrum of agreement and disagreement that exists between clinical experts, among themselves and with Flamingo-CXR and where there has been disparity, we have taken this as an opportunity to develop a collaborative assistive setting, with Flamingo-CXR and clinicians working together to improve clinical accuracy.

Results

The Flamingo-CXR report generation model is developed by fine-tuning the Flamingo vision–language foundation model⁸ on the task of generating a radiology report for a chest X-ray (CXR), using training data from two large deidentified datasets of CXR images and the corresponding radiology reports: (1) the MIMIC-CXR dataset²⁷, which is the largest public CXR dataset, acquired from a US emergency department, and (2) the INDI dataset²⁸, obtained from in/outpatient settings across India (see Methods and Extended Data Table 1 for further details of model training). To measure the quality of reports generated by our model, we conduct an expert radiologist evaluation of the generated reports, and we also use a set of report generation metrics, including two widely used clinical metrics: (1) the CheXpert F_1 score and (2) the RadGraph F_1 score, which measure the similarity between generated reports and original reports; we also use a set of widely adopted natural language generation (NLG) metrics (see Methods and Extended Data Table 1 for further details of model training).

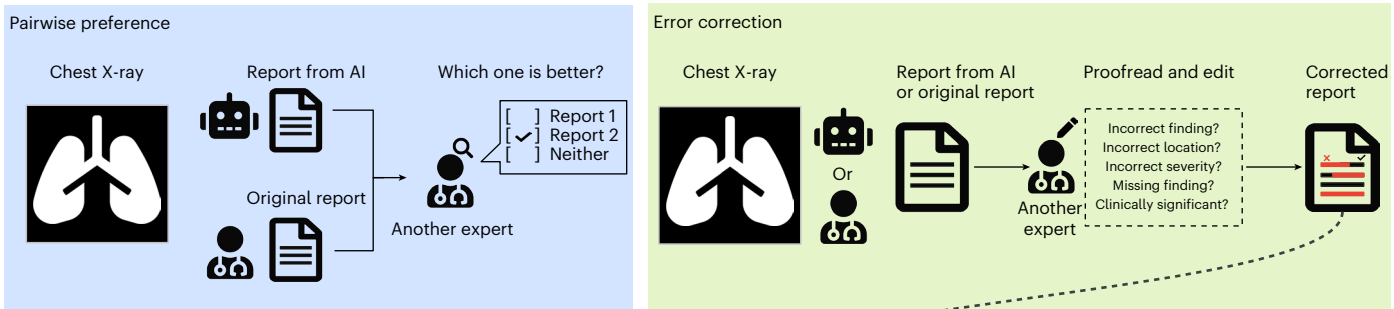
Automated report generation metrics

We find that Flamingo-CXR achieves a CheXpert F_1 score of 0.519 and a RadGraph F_1 score of 0.205 on the MIMIC-CXR dataset (Table 1). Among the methods capable of generating both the 'findings' and 'impression' section, Flamingo-CXR has outperformed the current state-of-the-art (SoTA) method by a large margin, attaining a 33% improvement relative to 0.389 as measured by the CheXpert F_1 score (R2GenGPT²⁹) and a 33% improvement from 0.154 as measured by the RadGraph F_1 score (CvT-21DistillGPT2 (ref. 13)) (see Methods for further details). For the sake of completeness, we also list CheXpert F_1 scores and RadGraph F_1 scores for models that only generate the 'findings' sections of reports. Even though our model is evaluated across a longer portion of text, the overall F_1 scores are still competitive, with a CheXpert F_1 score that is 1% greater than the current SoTA method even though this was evaluated on the Findings section alone (Med-PaLM-M²², 12B). In terms of the NLG metrics (CIDEr, BLEU4 and Rouge), the results are mixed; we achieve competitive BLEU4 and Rouge scores while attaining a compromised CIDEr score (Extended Data Table 2). This is also consistent with the established observation that NLG metrics do not reflect the clinical accuracy of the generated reports^{18,21,30}, for which our model, in particular, confers an improvement over the relevant previous methods.

Disease classification in comparison with human radiologists

For the INDI dataset, Fig. 2a shows that the generated reports of our model are overall as accurate (in terms of the microaveraged F_1 score) as one of the two radiologists in describing six clinical conditions in chest radiographs (namely, cardiomegaly, pleural effusion, lung opacity, edema, enlarged cardiomedastinum and fracture). For conditions that are frequent in the training dataset such as cardiomegaly and pleural effusion, we attain comparable or even superior agreement with the experts labels (as measured in the Kendall's tau coefficients) with respect to the two held-out radiologists (Fig. 2b). On the other hand, for under-represented conditions such as edema and enlarged

a Comparison with human experts



b Clinician + AI collaboration

Fig. 1 | Schematic overview of our human evaluation framework. a, To compare radiology reports generated by our AI model with reports written by human experts, we devise two evaluation schemes: (1) a pairwise preference test in which a certified expert is given two reports without knowing the source of the report (one report from our model and the original report from a radiologist) and they are asked to choose which report should be ‘used downstream for the care of this patient’; and (2) an error correction task in which a single report (either AI-generated or the original one) is evaluated carefully and edited if required. The expert is also asked to give the reason for each correction and to indicate

whether the error is clinically significant or not. **b,** We measure the utility of the AI-based report generation system in an assistive scenario in which the AI model first generates a report and the human expert revises as needed. For this task, we repeat the same pairwise preference test as before but this time the expert is asked to compare an AI-generated report corrected with human edits against a report written by human alone. We perform this evaluation on two datasets, one acquired in outpatient care delivery in India and another from intensive care in the United States. Board-certified radiologists are recruited in both countries to study the regional inter-rater variation.

Table 1 | Comparison of automatic report generation metrics on the MIMIC-CXR dataset

Model	Sections	Clinical metrics		
		CheXpert F_1 (all)	CheXpert F_1 (top 5)	Radiograph F_1
CXR-RePair ¹¹	Findings only	0.281	–	0.091
M2 Transformer ¹²	Findings only	–	0.567	0.220
RGRG ³⁹	Findings only	0.447	0.547	–
Med-PaLM-M ²² , 12B	Findings only	0.514	0.565	0.252
R2Gen ¹⁰	Findings+Impressions	0.228	0.346	0.134
WCT ¹⁴	Findings+Impressions	0.294	–	0.143
CvT-21DistillGPT2 (ref. 13)	Findings+Impressions	0.384	–	0.154
BioViL-T ¹⁵	Findings+Impressions	0.317	–	–
R2GenGPT ²⁹	Findings+Impressions	0.389	–	–
Flamingo-CXR (Ours)	Findings+Impressions	0.519	0.580	0.205

The clinical metrics for models that generate the ‘Findings’ sections (top) and the ‘Findings’ and ‘Impressions’ sections (bottom) for MIMIC-CXR radiographs are listed. Flamingo-CXR is trained to generate both ‘Findings’ and ‘Impressions’, and we observe that it outperforms the current SoTA method by 33%, when compared with other models that also generate ‘Findings’ and ‘Impressions’ sections. CheXpert F_1 (all) denotes the microaveraged F_1 score across all 14 categories of findings, whereas CheXpert F_1 (top 5) shows the same metric but over the most prevalent five categories from the MIMIC-CXR dataset (atelectasis, cardiomegaly, edema, consolidation and pleural effusion). All metrics are reported on the preprocessed test set ($n=1,931$). For all metrics, the higher the better, and the best results are shown in bold. An extended version of this table with NLG metrics is provided in Extended Data Table 2.

cardiomediastinum with extremely low prevalence rates (0.19% and 0.15%, respectively), the agreement scores of our model are lower than the two radiologists. The receiver operating characteristic (ROC) curves for the individual conditions (Extended Data Fig. 2) exhibit patterns consistent with such variation in the accuracy across conditions of different prevalence (see Methods for further details).

Expert evaluation of AI-generated and human-written reports
To achieve a more fine-grained and realistic assessment of the clinical quality of radiology reports generated by our model, we conduct an expert evaluation for reports in both the MIMIC-CXR and IND1 datasets. We recruit a group of 11 radiologists in the United States and 16 in India with board certification to perform two complementary evaluation

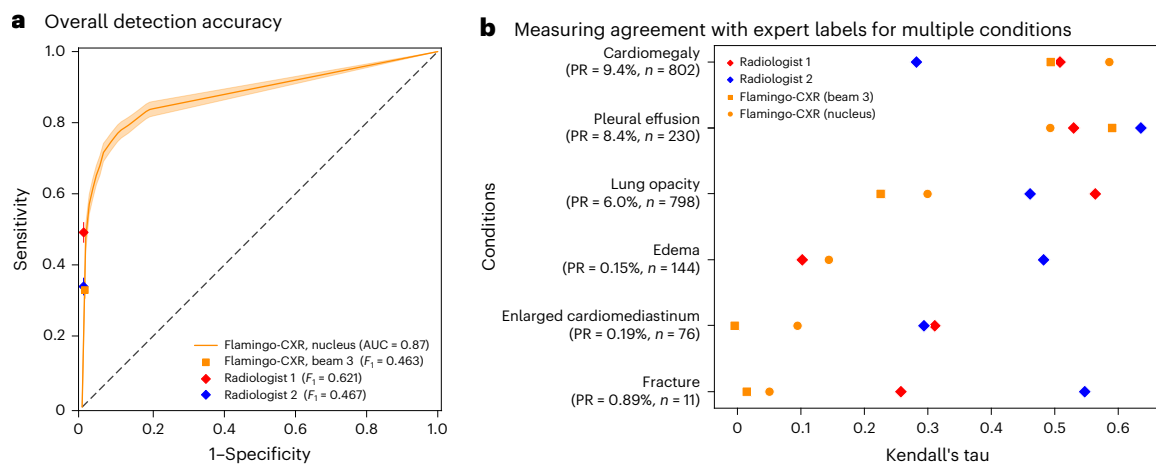


Fig. 2 | Comparison of detection accuracy with expert labels on the IND1 dataset. a, The ROC curve of the Flamingo-CXR report generation model with stochastic generation method (Nucleus) and corresponding area under the curve (AUC), shown along with the sensitivity and 1 – specificity pairs for two certified radiologists. The operating point of our model with the default deterministic inference scheme (Beam 3) is also shown. Details of the two inference algorithms are available in the Methods. The curve and the metrics are microaveraged across six conditions (cardiomegaly, pleural effusion, lung opacity, edema, enlarged cardiomeastinum and fracture) for which the labels were collected (n = 7,995 is the total number of IND1 test set reports). The GT labels are defined as the majority vote among the 5 labels obtained from the pool of 18 certified

radiologists. Error bars represent 95% confidence intervals (calculated using bootstrapping with 1,000 repetitions). **b**, Kendall's tau coefficients with respect to the expert labels are shown for the two held-out radiologists as well as for two inference schemes of our Flamingo-CXR model. We use the 'soft' labels derived by averaging over the available annotations instead of the majority vote labels as the target for computing the metric. On the vertical axis, the prevalence rates (PRs) of the respective conditions in the training set and their sample size in the test set are also shown. The target labels are the probabilities over the presence of the respective conditions calculated by averaging the binary condition labels from the expert pool.

tasks, namely (1) a pairwise preference test and (2) an error correction task (Fig. 1a and Extended Data Fig. 1; see Methods for further details).

Pairwise preference test. In this evaluation task, radiologists are provided with (1) a frontal view of a CXR image, (2) a radiology report generated by our AI system and (3) the original report written by a radiologist. They are asked to describe their preference from three options: report A, report B or equivalence between the two (that is, 'neither is better than the other'). Furthermore, they are asked to provide a justification for their preference, in free-form text, to better understand strengths and limitations of both reports (Extended Data Fig. 1a).

Across both datasets, generated reports from Flamingo-CXR were often considered preferable or equivalent to the ground truth (GT) report (Fig. 3 and Extended Data Fig. 3). For instance, in 77.7% of IND1 cases (and 56.1% of MIMIC-CXR cases), Flamingo-CXR reports were rated as equivalent or preferred relative to the original clinician report by at least half of the radiologists in our panel (Fig. 3a). Furthermore, in 94% of normal IND1 cases, Flamingo-CXR reports were rated as equivalent or preferred relative to the original clinician report by at least half of the radiologists in our panel (Fig. 3c). For this normal in/outpatient setting, more raters gave an equivalence rating rather than a preference rating for Flamingo-CXR reports (Extended Data Fig. 3), which is expected, given that normal in/outpatient reports have a relatively stereotypical structure that makes it difficult to discern differences between high-quality reports. In other settings, the majority of raters indicate a preference for Flamingo-CXR reports ahead of equivalence with original reports. Although these are strong results, it is clear that MIMIC-CXR reports are more challenging to model, which is not entirely surprising given that the MIMIC-CXR training dataset size is smaller and also contains a greater diversity of reports compared with the in/outpatient IND1 setting. To better understand the inter-rater diversity, we grouped all of our preference results according to the level of agreement between raters, from unanimity and majority to minority. This analysis reveals substantial disagreement among raters, who only reach unanimity (for Flamingo reports or GT reports) in

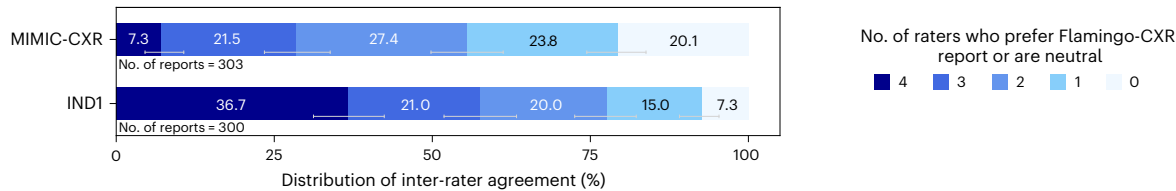
their preferences in 27.4% of MIMIC-CXR cases and 44% of IND1 cases. Across rater locations (India and the United States), the distribution of inter-rater variability is reasonably consistent (Fig. 3b). The strongest agreement is observed for normal IND1 cases, where 76% of cases reach agreement (with only 1% agreement for GT reports). By reporting progressive degrees of agreement and disagreement, our results can be interpreted relative to the desiderata of specific application scenarios, which may require greater or lesser degrees of agreement.

Last, in Fig. 3d, we provide a comparison of representative examples of AI-generated and human-written reports with varying degrees of inter-rater preference agreement. We also share the corresponding preference reasons from the respective raters. The top example shows a case for which the Flamingo-CXR report was preferred or rated as equivalent to the original clinician's report by all four radiologists on the panel. In this example, the raters explained that the Flamingo-CXR report correctly ruled out the 'retrocardiac opacity' originally noted, and also expressed caution against potential over-diagnosing in the original report of 'left lower lobe pneumonia/aspiration', recommending a repeat radiograph if clinically warranted (which is consistent with the conditional request for a repeat radiograph in the Flamingo-CXR report). We also give an example of a report in which all four radiologists prefer the clinician report and another where the panel is split 50:50.

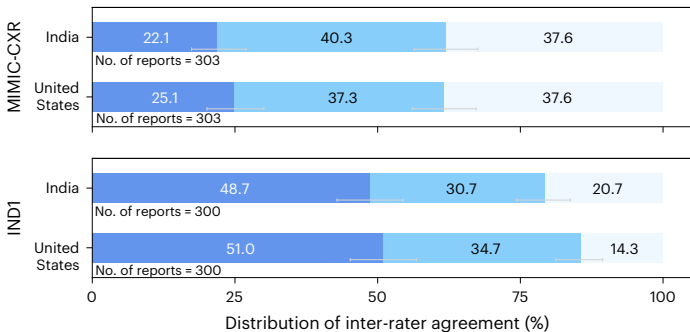
Error correction. In the error correction evaluation, the expert raters are provided with (1) the CXR image (a frontal view), and (2) a radiology report for this image, consisting of the findings and impression sections. Their task is to assess the accuracy of the given radiology report by identifying errors in the report and providing suggested replacements (Extended Data Fig. 1b).

Our results show that a non-negligible percentage (>10%) of the GT reports contain clinically significant disagreements for both MIMIC-CXR and IND1 datasets (upper row in Fig. 4a). The frequency of disagreement is also considerably different between the two locations of raters; Fig. 4b shows that the US-based radiologists disagree with the

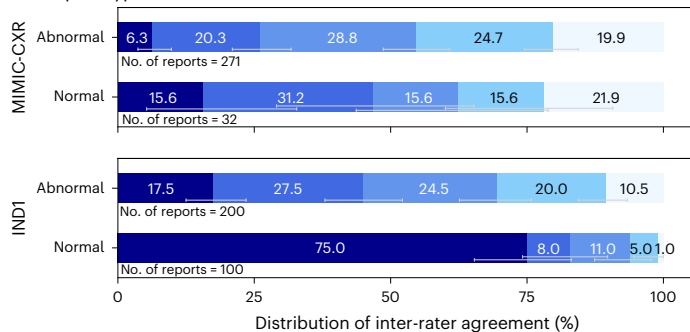
a Overall



b Rater locations



c Report type



d Examples

Image	Clinician's report	AI report	No. of votes for	Preference reasons
	<p>FINDINGS: Basilar opacity seen on the lateral view best corresponds to a retrocardiac opacity suspicious for developing left lower lobe pneumonia or aspiration event in the setting of altered mental status. Chronic peribronchovascular opacities seen bilaterally are similar in distribution and slightly more apparent due to lower lung volumes and AP technique. There is no pleural effusion or pneumothorax. The heart size is normal with normal cardiomeastinal silhouette. IMPRESSION: Left lower lobe opacity likely reflects pneumonia or aspiration.</p>	<p>FINDINGS: Lung volumes are low, accentuating the cardiac silhouette and bronchovascular structures. With this limitation in mind, there are no definite new areas of consolidation to suggest the presence of pneumonia. If clinical suspicion persists, repeat radiograph with improved inspiratory level may be helpful for more complete assessment of the lung bases.</p>	4/4	<p>For AI report:</p> <ul style="list-style-type: none">- No definite retrocardiac opacity identified, so I chose report B.- More accurate.- Findings are consistent with the radiograph.- I feel that labeling a left lower lobe pneumonia/aspiration is overenthusiastic in the limited setting and if there is a clinical suspicion then a repeat radiograph is warranted.
	<p>FINDINGS: Portable AP chest radiograph demonstrates worsening bilateral pleural effusions and associated atelectasis, greater on the right. There is also worsening pulmonary vascular congestion. There is no pneumothorax. Right internal jugular catheter probably terminates in the right atrium. IMPRESSION: Worsening pulmonary edema and bilateral pleural effusions. Results were relayed to ___. PA-C by phone at approximately 6:00 p.m. on ___.</p>	<p>FINDINGS: The patient is status post median sternotomy and CABG. A right internal jugular central venous catheter tip terminates at the junction of the SVC and right atrium. No pneumothorax is identified. Moderate to large layering bilateral pleural effusions are again demonstrated, slightly larger on the right and slightly improved on the left. Bibasilar airspace opacities likely reflect atelectasis. IMPRESSION: Right internal jugular central venous catheter tip at the junction of the SVC and right atrium. No pneumothorax. Moderate to large layering bilateral pleural effusions and bibasilar atelectasis.</p>	2/4	<p>For AI report:</p> <ul style="list-style-type: none">- [Radiologist's] Report does not mention post-sternotomy status.- Findings are consistent with radiograph. <p>For clinician's report:</p> <ul style="list-style-type: none">- Right IJ line likely terminates in the right atrium so I chose report X.- Pulmonary findings in reports are similar.- More concise. Both fail to mention small lung volumes and chin flexed significantly limiting exam.
	<p>FINDINGS: Low lung volumes are present. This accentuates the size of the cardiac silhouette which is likely mildly enlarged. Mediastinal and hilar contours are likely within normal limits. A right brachiocephalic venous stent is re-demonstrated. There is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. No pleural effusion or pneumothorax is identified. IMPRESSION: Low lung volumes with mild pulmonary vascular congestion.</p>	<p>FINDINGS: The heart is mildly enlarged. The mediastinal and hilar contours are unremarkable. There is no pleural effusion or pneumothorax. The lungs appear clear within the limitations of technique. IMPRESSION: No evidence of acute disease.</p>	0/4	<p>For clinician's report:</p> <ul style="list-style-type: none">- Report is more detailed and accurate.- Probable mild pulmonary vascular congestion as suggested in report / Also report mentions right brachiocephalic venous stent.- I agree with the report findings of mild vascular congestion.- Describes the positive findings in detail.

Fig. 3 | Results of pairwise preference test for MIMIC-CXR and IND1. a, Preferences for Flamingo-CXR reports relative to original clinician reports. Reports are grouped according to the level of agreement between reviewers. **b,** Clinician preferences for Flamingo-CXR reports depending on the location of the clinician, from either the US-based cohort or the India-based cohort. Note that there are two reviews from each location cohort, so in this case, unanimity corresponds to agreement between two clinicians rather than four in the full panel. **c,** Preferences for normal reports and separately, for abnormal reports. In

all panels, data are presented as mean values and error bars show 95% confidence intervals for the cumulative preference scores. **d,** Examples from MIMIC-CXR with varying degrees of inter-rater preference agreement; for two examples, all four radiologists unanimously preferred the AI report or the clinician's report, whereas for the remaining one, the preferences were divided equally. AP, anterior-posterior; CABG, coronary artery bypass graft; IJ, internal jugular; PA-C, physician assistant - certified; SVC, superior vena cava.

GT reports more often than the India-based radiologists. Last, we also observe from Fig. 4c that the GT reports for abnormal cases contain errors more often than the normal cases, likely caused by the higher variability and complexity of report contents.

The relative frequency of errors between the AI system and the human experts varies across the two datasets. Figure 4a (lower row) shows that, for the IND1 dataset, the model makes fewer errors (0.31) on average than the human experts (0.39), although the frequency of clinically significant errors is marginally higher (0.23 versus 0.20).

By contrast, for the MIMIC-CXR dataset, more (clinically significant) disagreements on average were reported in the AI-generated reports than in the original reports with a larger gap from 0.49 (0.28) to 0.27 (0.14) in terms of the average number of errors per report. Further decomposing this comparison into the distinct locations of raters in Fig. 4b reveals that the above patterns are largely preserved between the radiologists in the United States and those in India, but there remain a couple of noteworthy differences. The US-based raters reported considerably more disagreements on average than the India-based raters

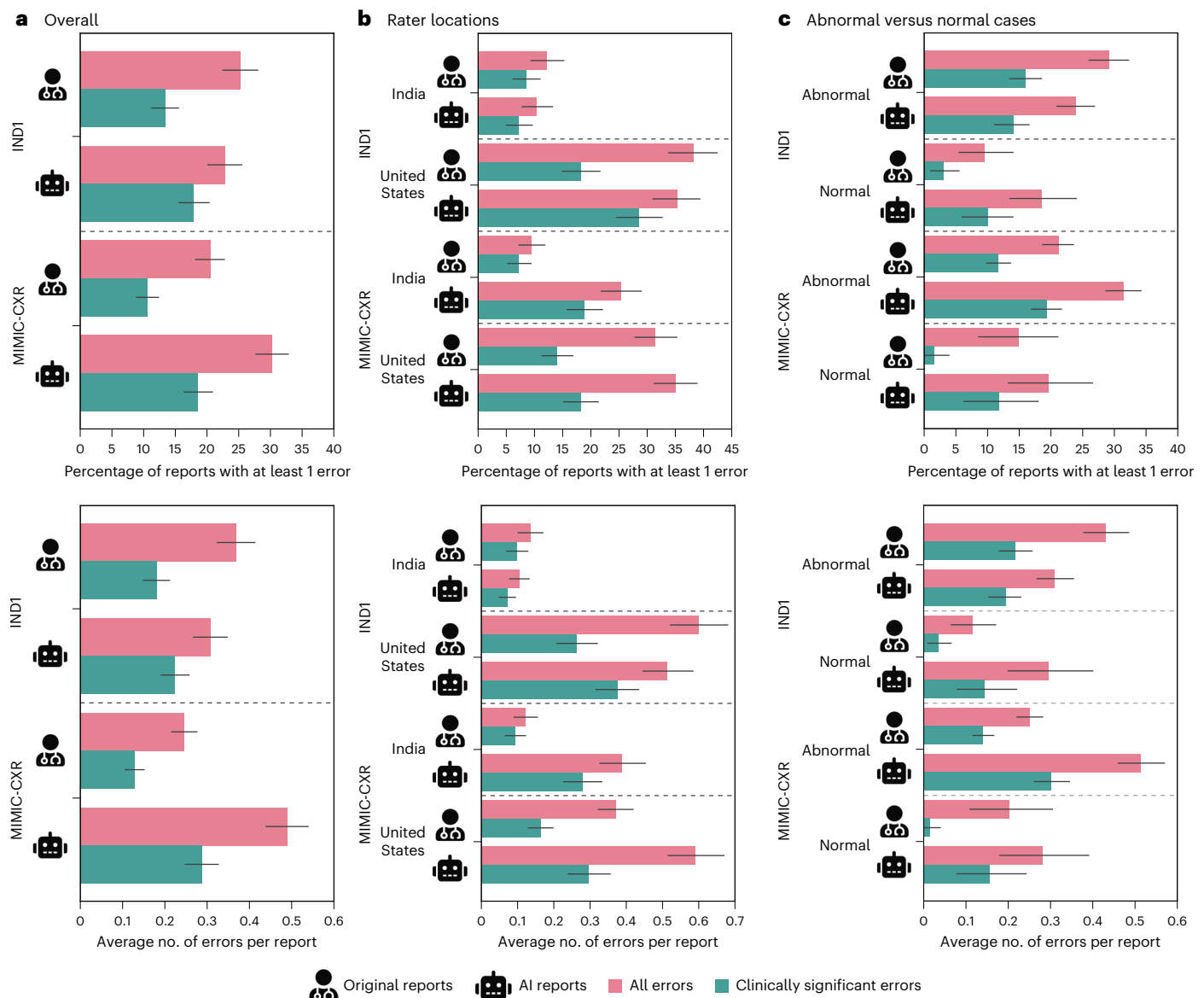


Fig. 4 | Comparison of error correction for the AI-generated reports and the original GT reports. a–c. The upper row shows the percentage of reports with at least one (clinically significant) error, and the bottom row shows the average number of identified (clinically significant) errors per report computed as the total number of detected errors divided by the number of all reports, including the ones without errors. These two metrics are compared across the IND1 and

MIMIC-CXR datasets overall (a), the two rater locations (India and the United States) to illustrate the regional inter-rater variation (b) and the normal and abnormal cases in the respective datasets (c). Error statistics for GT reports and Flamingo-CXR reports are given for each setting and grouped together as indicated by dashed lines. Data are presented as mean values and error bars correspond to 95% confidence intervals across cases and expert assessments.

across the board, particularly with more pronounced differences for IND1 dataset (acquired in India). It is known that there is a wide variety of radiology reporting styles, ranging from semi-structured free-form reports (for example, the MIMIC-CXR reports) through to a more structured style (for example, the IND1 reports) and these stylistic differences reflect the preferences of the clinicians who write those reports, the stylistic preferences taught by their radiology trainers along with their hospital and regional guidelines^{16,17}. These regional variations in reporting style are likely to account in part for observed regional variation in rater preferences. We also highlight that the raters in two locations are incongruent on the relative frequency of clinically significant errors for the IND1 dataset; the India-based raters flagged fewer errors in AI-generated reports than in the GT reports, whereas the reverse trend was observed for the US-based raters. Finally, Fig. 4c compares the amount of disagreement between the abnormal and

the normal cases. For the abnormal cases of IND1, marginally more clinically significant errors were reported in the human-written GT reports than in the AI-generated reports on average and vice versa for the MIMIC-CXR dataset.

To compare the distributions of error types across datasets, we explore the disagreement reasons for the edits made in reports (Extended Data Figs. 4a and 5). For both the model-generated reports and the original ones, the most dominant category of errors across the two datasets is the ‘incorrect finding’ category. The ‘incorrect finding’ category is less specific than the other two categories (‘incorrect severity’ and ‘incorrect location’). For the abnormal cases in the MIMIC-CXR dataset, statements with incorrect severity are much more common than those with incorrect locations in the original reports, whereas both are comparably frequent in the AI-generated reports. For the AI-generated reports (or human-written GT), 0.32 (0.14) errors on

average correspond to incorrect findings, 0.11 (0.03) are due to incorrect location of the finding and 0.09 (0.08) to incorrect severity. For the IND1 abnormal cases, however, the second most common error type is related to incorrect severity for both the GTs and AI reports. Overall, errors due to incorrect location of findings in the report (for example, opacity in left versus right lung) are more prevalent for the MIMIC-CXR abnormal cases than for the abnormal cases in IND1.

Lastly, in Extended Data Fig. 4b we show the differences and intersection of cases with errors between the original reports and the ones generated by our model. It is worth noting that for this analysis we consider (clinically) significant errors to be present in a case if at least one of the four raters identified an error in the corresponding report. Large proportions of the clinically significant errors are non-overlapping (72.7% for MIMIC-CXR and 59.7% for IND1 of total cases with at least one clinically significant error, respectively), suggesting frequent inconsistency in detected issues between the AI-generated reports and the original ones. Notably, in 27.3% and 22.7% of such cases in the MIMIC-CXR and IND1 datasets, clinically significant errors were identified only in the human reports, but not in the corresponding AI-generated reports. Some examples are provided in Extended Data Table 3 illustrating the nuanced nature of these differences. By contrast, there are also a considerable number of instances in which the AI-generated reports contain clinically significant errors, but the original reports do not. Examples of such instances are provided in Extended Data Table 4; some of these errors pertain to the limited spatial reasoning and counting capabilities of visual-language models. The presence of such disparities suggests that there may be potential complementarity between the AI system and the human experts in composing accurate radiology reports, which motivates us to investigate the utility of CXR-Flamingo in a clinician-AI collaboration setting.

Clinician-AI collaboration

In this section we explore collaboration between clinicians and Flamingo-CXR. For this collaboration, Flamingo-CXR produces a first draft report, and then a radiologist edits the report if necessary, by replacing sentences from the first draft with alternative sentences or by adding additional sentences to the report (Fig. 1b). The radiologists can make as many changes to the first draft report as they wish. We use the replacement sentences collected from the error correction task to produce these collaborative reports. To evaluate the quality of these clinician-AI reports, we ask our expert raters to indicate their preference for clinician-AI reports relative to the corresponding original clinician reports (Methods).

In Fig. 5d, we see an example of a clinician-AI report, in which a radiologist decided to replace sentences in the AI report that mentioned 'pneumothorax' with new sentences that mention hydropneumothorax instead. All four radiologists in our panel indicated that the clinician-AI report was preferable (or equivalent) to the original MIMIC-CXR clinician report, because the clinician-AI report was 'more succinct' and 'convey's the clinical findings better' [sic] and because of the statements concerning 'Right side pleural effusion and hydro-pneumothorax'. By contrast, for the AI report without edits, all four radiologists indicated a preference for the original clinician report because there was 'no residual pneumothorax' and because of the 'More accurate lung findings'.

For 53.6% of the MIMIC-CXR cases, we find that clinician-AI reports were rated as equivalent or preferred relative to the original clinician report, by at least half of the radiologists in our panel (Fig. 5a). In comparison, for reports generated by Flamingo-CXR alone without collaboration, 44.4% of reports were rated as equivalent or preferred relative to the original clinician report, by at least half or more of the radiologists in our panel. We observe similar findings for IND1, where the reports from the clinician-AI collaboration were rated as preferable or equivalent by half or more of the radiologists in 71.2% of cases,

in comparison with 51.2% for reports generated by Flamingo-CXR alone. We also observe variation in the preference results between normal and abnormal reports, and between different cohorts of collaborating clinicians, most likely reflecting variations in stylistic preferences across regions (Fig. 5b and Extended Data Fig. 6).

Discussion

In this work, we present Flamingo-CXR, a state-of-the-art AI radiology report generation system for chest radiographs built by specializing a recent vision-language foundation model⁸ on this challenging task. Our model achieves competitive performance in multiple automated metrics in two clinical contexts and geographical locations, namely intensive care in the United States and in/outpatient care delivery in India. To gauge the clinical quality and potential real-world utility of our report generation system we perform the most comprehensive expert evaluation of AI-generated reports published to date, and compare these with human-written GT reports with a group of certified radiologists. This evaluation is performed both in an autonomous and an assistive AI context. In addition, nuanced feedback from clinicians provides insight into disparities and defines areas for future enhancement.

Previous work has repeatedly reported the shortcomings of automated 'natural language generation' metrics for assessing reports of radiology images²¹. However, the majority of published works on the development of AI systems for this task, including recent approaches with acclaimed state-of-the-art performance, solely report automated metrics, while the direct proximity to expert accuracy and potential clinical utility remains unknown. Only a handful of previous works have attempted to evaluate AI systems with human experts. We go further in this work, in our fine-grained exploration of diversity and granularity of expert radiologist evaluations. For example, a similar evaluation schema for the same US dataset (MIMIC-CXR) was previously explored²², but assumed that the GT report is correct, without evaluating the inter-rater variability inherent in chest radiograph interpretation³¹. In another recent study, AI-generated reports for in-house emergency chest radiographs were compared against experts, revealing that the quality, on average, was only marginally inferior to that of on-site radiologists and surpassed that of teleradiology reports²³. However, both studies only evaluated the AI report generation model as a stand-alone system on a dataset acquired in an emergency department in the United States, whereas our study considers a more diverse setup that encompasses both autonomous and assistive scenarios for datasets from intensive care in the United States as well as in/outpatient care delivery in India, using evaluations from two distinct groups of clinicians, working in India and in the United States. Furthermore, our study enriches this evaluation by collecting granular information on error types (for example, distinction between incorrect findings, location and severity), and provides fine-grained insights into how the AI system differs from human experts, which was absent in the previous works.

Human evaluation results shed more light on the aspects of our model's report quality that might inform and enable applications of the technology in future clinical workflows. Notably, for the normal IND1 cases, the raters unanimously viewed the AI-generated reports to be at least equivalent to the human reports in 75% of the cases. This strong performance on normal cases suggests potential clinical applicability in using the report generation model in the subset of such in/outpatient cases (for instance, taken alongside previous works that show AI systems to have strong accuracy in predicting whether CXRs are normal or abnormal²⁸), allowing radiologist attention to be allocated to patients with abnormalities. However, we notice there is considerable room for improvement for MIMIC-CXR whose original reports are in general more detailed and less templated than IND1.

This inter-dataset discrepancy in report quality highlights the importance of evaluation in different clinical contexts and geographic

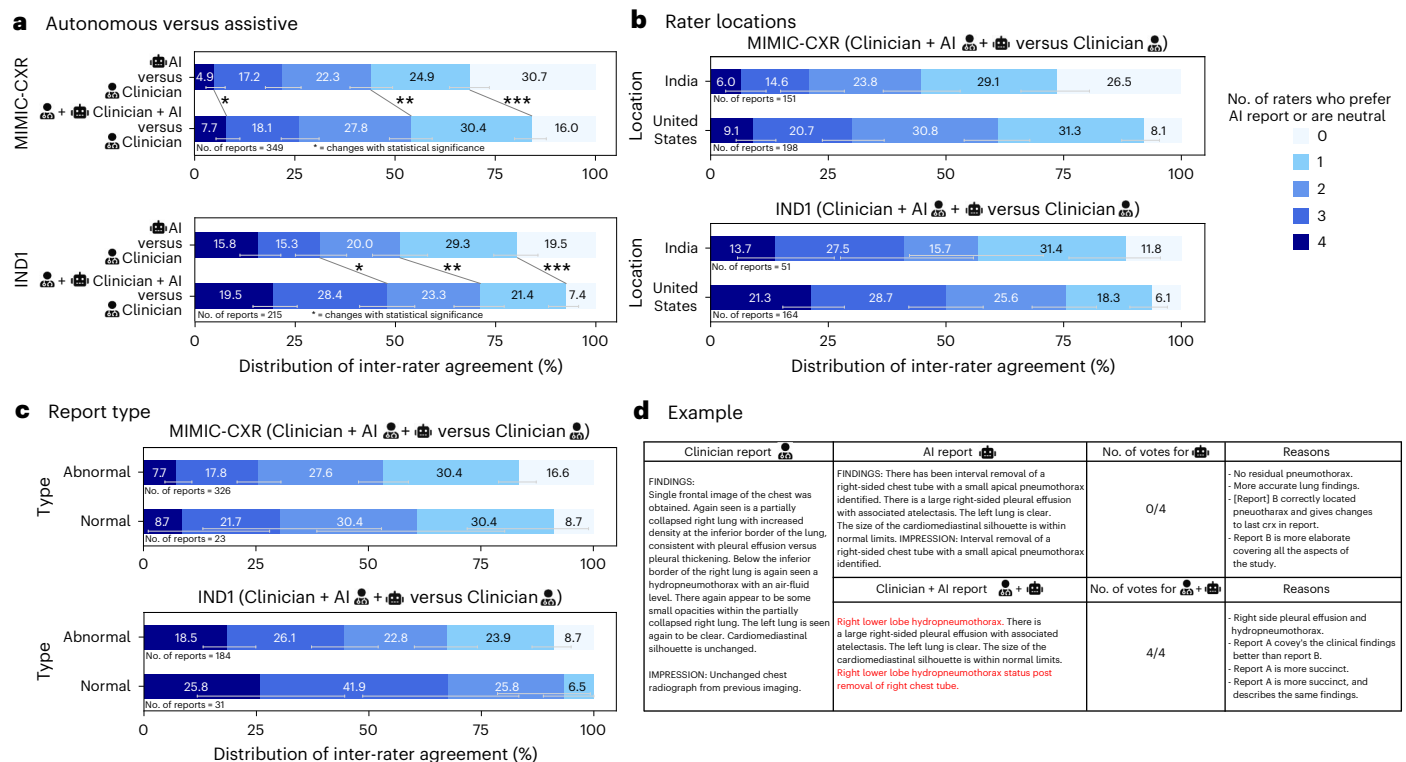


Fig. 5 | Results of pairwise preference test for clinician–AI collaboration. **a**, Preferences for reports produced from the clinician–AI collaboration relative to the original clinicians’ reports are shown here. The corresponding preference scores for reports produced by Flamingo-CXR without human collaboration are also given. Reports are grouped by the level of agreement between reviewers, and in all cases, we show results for the subset of reports that required editing during the error correction task. Data for all panels are presented as mean values and error bars show 95% confidence intervals for the cumulative preference scores. Significant differences ($P < 0.05$) between clinician–AI results and AI-only results calculated using a one-sided chi-squared test are indicated by an asterisk (with MIMIC-CXR P values given by $*P = 1.3 \times 10^{-2}$, $**P = 5.7 \times 10^{-4}$, $***P = 3.2 \times 10^{-9}$;

and IND1 P values given by $*P = 1.2 \times 10^{-7}$, $**P = 4.4 \times 10^{-9}$, $***P = 7.7 \times 10^{-6}$). **b**, Preferences for reports produced from a collaboration between Flamingo-CXR and radiologists from our US-based cohort and separately, from our India-based cohort. **c**, Preferences for normal reports and separately, for abnormal reports. **d**, An example of a pairwise preference test for a clinician–AI report and an AI report, relative to the original clinician’s MIMIC-CXR report. All four radiologists initially indicated a preference for the original clinician’s report to the AI report. Another radiologist revised two sentences in the AI report (indicated in red), resulting in a complete flip in preference in which all four radiologists unanimously expressed the superiority (or equivalence) of the clinician–AI report.

regions, which was previously not considered. The desired contents of a report are ultimately contingent on the given clinical context, and assuming access to large quantities of training data from every plausible scenario is not realistic. Future work will consider reinforcing our system with the capability to follow user instructions^{28,32} so the users can control the outputs more flexibly through natural language and the capability to learn efficiently from a small quantity of data through techniques such as in-context learning³³ or parameter-efficient optimization³².

The complexity in evaluating the quality of radiology reports is underscored by the observed high inter-rater variability, as evidenced by: (1) identified (clinically significant) errors in the GT reports as part of the error correction task, and (2) the variability in both human evaluation tasks in terms of preferences and disagreements with report statements. For instance, there is unanimous agreement among our panel of raters in only 27.4% of MIMIC-CXR cases and 44% IND1 cases, respectively. This indicates the importance of our approach to obtaining multiple readings per case, unlike previous works that have only evaluated each case once²².

In-depth analysis shows that both human and AI systems can make errors in different ways, hinting at potential complementary properties between the two. Manual inspection unveils some examples in which nuanced clinical errors were detected in the human reports, but not in the corresponding AI-generated reports and vice versa (Methods and Extended Data Tables 3 and 4). Finally, another difference

between clinicians and our AI system is the input information at disposal when writing the reports. Integrating such extra information into our AI system will likely enhance the reporting accuracy¹⁵ but requires further study.

Moving beyond the autonomous setting, this work evaluates CXR report generation in an assistive setting. Our results indicate that AI-generated reports with expert revisions were reported to be preferable or equivalent to original clinician reports in 71.2% of IND1 cases in comparison with 51.2% of cases without expert revisions, and similarly, in 53.6% of MIMIC-CXR cases in comparison with 44.4% of cases without expert revisions, according to half or more of our raters. Our proof-of-concept evaluation exhibits the initial promise of AI report generation as an assistive system that augments the report writing process of radiologists.

These results are not without limitations. We have demonstrated the ability of Flamingo-CXR to generalize to previously unseen X-ray images from an intensive care setting (given by the standard MIMIC-CXR test set) and to an in/outpatient setting in India (given by the IND1 test set), but for other clinical settings that involve different types of data, such as CXRs with lateral views or other non-frontal views, CXRs from multiple time points and CXRs containing out-of-distribution conditions that do not appear in the training data, we expect that additional training data will be required for further fine-tuning our model. We also observe that the AI reports with human edits do not reach perfect preference or equivalence compared with

the original reports. There are several possible reasons for this. First, there is a baseline level of inter-rater variability both in the preference decision and the error correction process. Second, the location of the clinician making edits in the assistive setting has an impact on the preference decisions. This may reflect a difference in stylistic preferences across regions. We also observed some variability in the quality of edits (a whole sentence replaced with a single word; for example, ‘cardiomegaly’), which render the resultant reports quite unnatural despite being clinically correct. Third, it is possible that a clinician working in collaboration with AI may produce a report that is less accurate than a clinician working alone. Indeed, this is a common phenomenon observed in multiple lines of work in CXR classification tasks, where collaboration often result in less accurate predictions³. Clinician–AI collaboration typically becomes unhelpful when the experts overly rely on the AI predictions^{34,35} or are unduly critical of them³⁶. Development of strategies for identifying when to provide AI-generated reports is likely to be helpful for maximizing the benefits of AI assistance³⁷. Fourth, although it is plausible that revising an AI-generated report may require less time than composing a report from scratch, this work does not assess this explicitly and it is beyond the scope of the current work. Quantifying the time-saving aspect, however, warrants another carefully designed human study focused on measuring the reporting time of human experts, which commonly varies between individuals and is influenced by a plethora of factors such as the clinical context, reporting style, expertise and complexity of cases. Finally, clinician–AI collaborations can take more complex forms than our design and ideally should ultimately be bidirectional and interactive, much like an experienced colleague that answers the radiologist’s questions and provides high-quality feedback on their reports (for example, flagging potential errors and missing findings). Although we have witnessed initial signs of such possibilities in the recent work on interactive, multimodal medical AI^{26,33,38}, there remains a considerable amount of progress to be made toward building a clinically useful writing assistant for radiology.

Overall, our observation of a positive effect from clinician–AI teamwork is very encouraging, especially given the limitations outlined above, the possibilities for future developments and the clinical relevance of this setting, where most AI tools that are approved for clinical decision-making are deployed in an assistive rather than autonomous setting^{24,25}. Furthermore, our observation of strong baseline preference ratings for Flamingo-CXR reports without clinician assistance, especially for normal in/outpatient reports, is intriguing, and may already raise the possibility for potential clinical applicability. Finally, by moving beyond automatic evaluation metrics, by engaging expert clinicians for evaluations and error correction, across a diversity of regions, clinical settings and data types, we have been able to richly characterize the wide spectrum of agreement and disagreement that exists between clinical experts, among themselves and with Flamingo-CXR, and where there has been prevailing disparity, we have embraced this as an opportunity for collaboration between Flamingo-CXR and clinicians working together in an assistive setting. Although there are immediate possibilities for enhancements and applications, Flamingo-CXR is intended as an experimental research-only model, and not as a tool for clinical deployment. However, we hope that this work will encourage and support the wider research community to further explore the full nuance, complexity and variability of the socio-technical landscape induced by the application of visual–language models in radiology report generation and beyond.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-024-03302-1>.

References

1. Maru, D. S.-R. et al. Turning a blind eye: the mobilization of radiology services in resource-poor regions. *Global Health* **6**, 18 (2010).
2. Rimmer, A. Radiologist shortage leaves patient care at risk, warns Royal College. *BMJ* **359**, j4683 (2017).
3. Rajpurkar, P. & Lungren, M. P. The current and future state of AI interpretation of medical images. *N. Engl. J. Med.* **388**, 1981–1990 (2023).
4. Allen, B., Agarwal, S., Coombs, L., Wald, C. & Dreyer, K. 2020 ACR Data Science Institute artificial intelligence survey. *J. Am. Coll. Radiol.* **18**, 1153–1159 (2021).
5. Milam, M. E. & Koo, C. W. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. *Clin. Radiol.* **78**, 115–122 (2023).
6. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2018).
7. Guo, W., Wang, J. & Wang, S. Deep multimodal representation learning: a survey. *IEEE Access* **7**, 63373–63394 (2019).
8. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
9. Li, C. et al. Multimodal foundation models: from specialists to general-purpose assistants. *Found. Trends Comput. Graph. Vis.* **16**, 1–214 (2023).
10. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1439–1449 (eds Webber, B. et al.) (Association for Computational Linguistics, 2020).
11. Endo, M. et al. Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model. *Proc. Mach. Learn. Res.* **158**, 209–219 (2021).
12. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C. & Jurafsky, D. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 5288–5304 (Association for Computational Linguistics, 2021).
13. Nicolson, A., Dowling, J. & Koopman, B. Improving chest X-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **144**, 102633 (2023).
14. Yan, B. et al. Style-aware radiology report generation with RadGraph and few-shot prompting. *Empir. Method Nat. Lang. Process.* <https://doi.org/10.18653/v1/2023.findings-emnlp.977> (2023).
15. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision–language processing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15016–15027 (2023).
16. Hartung, M. P., Bickle, I. C., Gaillard, F. & Kanne, J. P. How to create a great radiology report. *Radiographics* **40**, 1658–1670 (2020).
17. Kahn, C. E. Jr et al. Toward best practices in radiology reporting. *Radiology* **252**, 852–856 (2009).
18. Liu, G. et al. Clinically accurate chest X-ray report generation. *Proceedings of the Machine Learning for Healthcare Conference. Proc. Mach. Learn. Res.* **106**, 249–269 (2019).
19. Jain, S. et al. RadGraph: extracting clinical entities and relations from radiology reports (version 1.0.0). *PhysioNet* <https://doi.org/10.13026/HM87-5P47> (2021).
20. Khanna, S. et al. RadGraph2: modeling disease progression in radiology reports via hierarchical information extraction. Preprint at <https://doi.org/10.48550/arXiv.2308.05046> (2023).
21. Yu, F. et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns (N Y)* **4**, 100802 (2023).

22. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* <https://doi.org/10.1056/Aloa2300138> (2024).
23. Huang, J. et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA Netw. Open* **6**, e2336100 (2023).
24. Harvey, H. B. & Gowda, V. How the FDA regulates AI. *Acad. Radiol.* **27**, 58–61 (2020).
25. Norden, J. G. & Shah, N. R. What AI in health care can learn from the long road to autonomous vehicles. *NEJM Catalyst* **3**, (2022).
26. Li, C. et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In *Proc. 37th Int. Conf. Neural Information Processing Systems* (Curran Associates Inc., 2024).
27. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
28. Nabulsi, Z. et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Sci. Rep.* **11**, 15523 (2021).
29. Wang, Z., Liu, L., Wang, L. & Zhou, L. R2GenGPT: radiology report generation with frozen LLMs. Preprint at <https://arxiv.org/abs/2309.09812> (2023).
30. Boag, W. et al. Baselines for chest X-ray report generation. In *Proc. Machine Learning for Health NeurIPS Workshop* Vol. 116 (eds Dalca, A. V. et al.) 126–140 (PMLR, 2020).
31. Gefter, W.B., Post, B.A. & Hatabu, H. Special features commonly missed findings on chest radiographs: causes and consequences. *Chest* **163**, 650–661 (2022).
32. Singhal, K. et al. Towards expert-level medical question answering with large language models. Preprint at <https://arxiv.org/abs/2305.09617> (2023).
33. Moor, M. et al. Med-Flamingo: a multimodal medical few-shot learner. In *Proc. 3rd Machine Learning for Health Symposium*, *PMLR* **225**: 353–367 (2023).
34. Rajpurkar, P. et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest X-rays in patients with HIV. *NPJ Digital Med.* **3**, 115 (2020).
35. Seah, J. C. Y. et al. Effect of a comprehensive deep-learning model on the accuracy of chest X-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digital Health* **3**, e496–e506 (2021).
36. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology* (National Bureau of Economic Research Inc., 2023).
37. Dvijotham, K. et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Med.* **29**, 1814–1820 (2023).
38. Chen, Z. et al. CheXagent: towards a foundation model for chest X-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models* (AAAI, 2024).
39. Tanida, T., Müller, P., Kaissis, G. & Rueckert, D. Interactive and explainable region-guided radiology report generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7433–7442 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

¹Google DeepMind, London, UK. ²Google Research, London, UK. ³Apollo Radiology International, Hyderabad, India. ⁴Present address: Open AI, San Francisco, CA, USA. ⁵Present address: GlaxoSmithKline AI, London, UK. ⁶These authors contributed equally: Ryutaro Tanno, David G. T. Barrett. ✉e-mail: rtanno@google.com; barrettdavid@google.com; alankarthis@google.com; iraktena@google.com

Methods

Ethical approval

The use of deidentified retrospective datasets was reviewed by Advarra IRB (Columbia, MD), which determined that it was exempt from further review under 45 CFR 46. The involvement of clinicians in this study, using the same deidentified retrospective data is also covered in this waiver.

Model

Our report generation model is built by fine-tuning a state-of-the-art vision–language foundation model, Flamingo⁸, which has attained impressive performance on data-efficient adaptation to new tasks. We fine-tune this model on the radiology report generation task, with an effective combination of regularization and adaptation techniques. Flamingo has a flexible transformer-based multimodal sequence-to-sequence architecture that can learn to integrate a mixture of medical images and reports with no model modifications.

Task

Our model is trained to generate both the ‘findings’ and ‘impression’ sections of the report for a frontal view (anterior–posterior or posterior–anterior) of the chest radiograph, which typically captures all the relevant observations the radiologist makes in a study. The model is not provided with additional projections, such as lateral views or prior views, other clinical history data or indication data. Flamingo-CXR only had access to the current radiograph at a lower resolution of 1 megapixel (in contrast with the original resolution of approximately 4 megapixels), whereas the original radiologists additionally had access to contextual information, patient history and previous scans. In the clinical setting, additional data, such as lateral views and prior views are often required, and we expect that fine-tuning our model with this data would enhance the capabilities of our model. However, recent studies do not use this additional data, so in our task formulation, we have also adopted this convention, which allows us to make a fair comparison with previously published benchmarks^{10,15,29,40}.

Architecture. Flamingo is a general-purpose family of transformer-based visual–language models that take visual data as input (for example, images), interleaved with text and produce free-form text as output. The key architectural components are (1) the language model that operates on the input text and generates the output text, (2) the vision encoder that maps visual data into the same representation space as text input and (3) the connective module that integrates both modalities. The combination of the perceiver resampler⁴¹ and cross-attention layers in this connective component offer an expressive way for the language model to incorporate visual information for the next-token prediction task. There are multiple versions of Flamingo at different scales, and our report generation model, Flamingo-CXR is built using a parsimonious 400 million parameter version. Flamingo models the likelihood of the radiology report y conditioned on the input image x in an auto-regressive fashion:

$$p(y|x) = \prod_{\ell=1}^L p(y_{\ell} | y_{<\ell}, x_{\leq \ell}),$$

where y_{ℓ} is the ℓ -th language token of the input report, $y_{<\ell}$ is the set of preceding tokens and p is parameterized by the model.

Optimization. We take a version of Flamingo, pretrained on a large set of interleaved text–image data, and fine-tune it on the specific task of radiology report generation by minimizing a weighted sum of the expected negative log-likelihoods of report given the chest radiograph over both MIMIC-CXR (United States) and IND1 (India) datasets:

$$\lambda_{\text{US}} \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{US}}} \left[- \sum_{\ell=1}^L w(x,y) \log p(y_{\ell} | y_{<\ell}, x_{\leq \ell}) \right] + \lambda_{\text{India}} \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{India}}} \left[- \sum_{\ell=1}^L w(x,y) \log p(y_{\ell} | y_{<\ell}, x_{\leq \ell}) \right],$$

where \mathcal{D}_{US} and $\mathcal{D}_{\text{India}}$ denote the MIMIC-CXR and IND1 datasets respectively, λ_{US} and λ_{India} are the data-specific coefficients that are tuned to maximize the benefits of jointly training on both datasets, and lastly $w(x,y)$ is a reweighting function that changes the amount of penalty depending on whether the example (x,y) contains any thoracic abnormalities. Specifically, we use importance weighting here⁴¹ and define $w(x,y)$ to output the inverse of the proportion of healthy cases in the corresponding dataset (if the given example is normal) or otherwise that of abnormal cases. This ensures that the model is equally penalized to compose inaccurate reports across the healthy and the abnormal cases; this is particularly important for the IND1 dataset in which the healthy cases account for more than 90% of the training data. We set the weighting coefficients $\lambda_{\text{US}} = 1.0$ and $\lambda_{\text{India}} = 0.5$.

To further enhance the reporting accuracy on abnormal cases, we augment the above training objective with an auxiliary classification loss for abnormality classification. To this end, we applied a published labeling software, CheXpert⁴² to extract the presence of multiple thoracic conditions from the training reports, derived binary abnormality labels (1 if any of the conditions is present or else 0), and used them to compute this auxiliary classification loss. We found the addition of this abnormality classification task to be helpful in improving the sensitivity of the generated reports across these conditions.

We optimize parameters using AdamW⁴³ with initial learning rate of 10^{-3} and $\beta = [0.9, 0.999]$ with batch size of 16 examples and we train for 150,000 steps. The above hyper-parameters are selected based on the overall microaveraged F_1 score for detection of CheXpert conditions on the validation set. The best checkpoint was selected based on the overall CIDEr-d score on the validation set. We freeze the language component and only update the parameters in the vision encoder and the connective component (perceiver resampler and cross-attention layers) because our initial experiments showed updating the language part resulted in overfitting and fine-tuning the rest of the architecture was important for adapting to the unfamiliar medical domain not represented in the pretraining datasets.

Inference. Once Flamingo is trained, we use it to generate the radiology reports on the test chest radiographs with two decoding strategies: beam search with the width size set to 3 and nucleus sampling⁴⁴ with $P = 0.9$. We used the former deterministic decoding method by default, and the generated reports are used in calculating of reported NLG and clinical metrics in Table 1 and Extended Data Table 2 as well as in the subsequent expert evaluation. However, we also used the latter stochastic decoding method when we needed to generate multiple reports. For example, to plot the ROC curves in Fig. 2 and Extended Data Fig. 2 for measuring the disease classification accuracy of reports, we used the nucleus sampling to generate 250 candidate reports, derived the condition labels from each with the CheXpert labeler and aggregated them to compute the per-condition probability.

Datasets and preprocessing

We developed and evaluated our automatic report generation model using two large deidentified datasets of CXR images and corresponding radiology reports from the United States and India. Chest radiography offers a valuable testbed for automatic report generation systems because it is the most widely used thoracic imaging modality in the world²⁸. Even for such a specific domain, the contents of radiology reports differ widely between geographic regions and clinical contexts. To account for these variations, we used the combination of the MIMIC-CXR dataset²⁷, acquired in the emergency department of the

Beth Israel Deaconess Medical Center in the United States, and another private research dataset of a similar scale, which we refer to as IND1 (ref. 28), obtained from a large hospital group in India. These datasets do not contain sex or gender information.

IND1. This is a deidentified dataset⁴⁵ of 263,021 frontal chest radiographs (digital and scanned) with reports obtained from five regional centers across a large hospital group in India (Bangalore, Bhubaneswar, Chennai, Hyderabad and New Delhi) between November 2010 and January 2018. We use the same training, validation and test split as in previous studies²⁸. Thus, a total of 250,066 samples are used for training, 4,960 samples for validation and 7,995 samples for testing of Flamingo-CXR. Furthermore, a small subset of 2,306 cases are annotated with varying numbers of binary labels (0, absent; 1, present) for six thoracic conditions (cardiomegaly, pleural effusion, lung opacity, edema, enlarged cardiomeastinum and fracture) obtained from a pool of 18 certified radiologists in the United States. The agreement labels are derived by calculating the majority vote, and used as the reference labels for evaluation of report quality in classification accuracy (for example, ROC curves in Extended Data Fig. 2 and F_1 scores in Extended Data Table 2).

MIMIC-CXR. As the largest public dataset to date, MIMIC-CXR²⁷ contains 377,110 images and 227,835 reports. In our experiments, we use the official split provided by the dataset resulting in 222,758 training examples, 1,808 validation examples and 3,269 test examples. For the reports, we remove redundant whitespaces (line breaks and so on). We only use frontal view scans (anterior–posterior and posterior–anterior views) and discard samples where only lateral views are provided. We only keep the FINDINGS and IMPRESSION sections of reports and filter out cases that do not contain an IMPRESSION section, following previous studies¹⁵.

Lastly, more than 50% of the examples in MIMIC-CXR contain previous scans¹⁵ and the corresponding reports often describe findings in reference to these measurements (see the highlighted sentence in the left column of Extended Data Table 1 for an example). Consequently, as also reported in recent work⁴⁶, naively training on the entirety of the MIMIC-CXR data leads to a model that generates reports with hallucinated references to nonexistent previous reports (see the right column; note that the model only has access to the current radiograph). To ameliorate this issue, we remove all the training examples with references to previous studies (see the middle column for an example of the improved prediction as a result). However, we still report the evaluation metrics on all the test examples for a fair comparison with the previous studies. The combination of all the above preprocessing and filtering steps result in 90,968 training, 688 validation and 1,931 test examples.

Image processing. All images in both datasets are resized to 320×320 while preserving the original aspect ratio, padded if needed, and normalized to zero mean and unit standard deviation. Color jitter and resize/crop transformations are applied as data-augmentation during the training of Flamingo-CXR.

Icons in Figs. 1, 3, 4 and 5 and Extended Data Figs. 4 and 6 were sourced from Font Awesome (<https://fontawesome.com>) under the CC BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>).

Automated report generation metrics

We report performance on established automated metrics to facilitate comparison with previous studies, using two different categories of metrics. The first category are the NLG metrics that include the CIDEr score⁴⁷, BLEU score⁴⁸ and Rouge-L^{47,49}, which are widely used measures of report quality. However, multiple studies^{18,30,50,51} have recently highlighted the inadequacy of these NLG metrics for assessing factual correctness and consistency, key properties for determining the clinical utility and quality of radiology reports.

We also compute another category of metrics that are specifically designed to measure the accuracy of descriptions for relevant clinical findings, and we refer to them as clinical metrics. Specifically, following previous work^{10,12,15,18}, we report the microaverage F_1 score across 14 distinct categories related to thoracic diseases and support devices (atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomeastinum, fracture, lung lesion, lung opacity, no finding, pleural effusion, pleural other, pneumonia, pneumothorax and support devices). To ensure a fair comparison with previous publications on the MIMIC-CXR dataset, we use the CheXpert labeling software⁴², to extract from the reports the binary labels that indicate the presence of these radiological findings. We refer to this metric as CheXpert F_1 . For the IND1 dataset, published results on classification performance are unavailable, so we use labels for these findings that were collected in a separate study⁴⁵ from a group of 18 board-certified radiologists (American Board of Radiology) in the United States, and we use the corresponding consensus labels as GTs. In this way, we aim to mitigate the known inaccuracy of the CheXpert labeler software and have a test set with a more reliable metric of clinical factual correctness. Finally, to align with more recent studies^{21,22}, we also report the RadGraph score^{49,20}, which not only accounts for the presence of these findings but also accounts for the relationships between them and other image features (for example, anatomical locations). All these results are reported on held-out test data that was not used to train or tune the model.

Disease classification in comparison with human radiologists

In Fig. 2 and Extended Data Fig. 2, the GT labels are derived from the majority votes of five annotations per example acquired by a separate group of 18 experts and, thus, should provide more reliable labels than the ones extracted from the CheXpert labeler⁴² (which was used for the MIMIC-CXR dataset). To generate the binary labels from the generated reports from Flamingo-CXR, the CheXpert labeler is used as before.

Expert evaluation of AI-generated and human-written reports

An accumulation of evidence has shown that automatic report generation metrics fail to appropriately evaluate many nuanced issues in radiology reports²¹. Here we describe how we evaluate AI-generated reports by conducting radiologist evaluation tasks. To document human errors in report writing and to characterize differences in quality with our AI system, we also evaluate the original reports (that we have treated as GTs) by obtaining additional readings from different radiologists than the ones who provided the original reports.

Annotators. We recruited a group of 16 radiologists in India and 11 radiologists in the United States with board certifications (Diplomate of National Board and American Board of Radiology, respectively). All raters performed the required Collaborative Institutional Training Initiative (CITI) training before performing the evaluation tasks on the MIMIC-CXR dataset. None of the raters were coauthors of this work and the raters were not given any information about the origin of the reports, including the possibility that the reports may be generated by an AI model. We ask four radiologists to evaluate each report, two from the US cohort and two from the India cohort. This allows us to represent inter-rater preference variability and regional preference variability. We highlight that radiologists that provided annotations for the first phase of error correction or preference test tasks were excluded from the human–AI collaboration evaluation to avoid annotation bias. Before the large-scale evaluation, we validated the labeling interface with an expert to ensure that instructions were clear and opt-out options were available where essential.

Sample selection. We randomly select a fixed number of normal and abnormal cases from the IND1 and MIMIC-CXR datasets. To ensure good coverage of different abnormalities the set of abnormal cases reviewed by radiologists was larger than the one for normal cases. In total, 606 cases were evaluated by expert radiologists in the two tasks:

34 normal and 272 abnormal cases from the MIMIC-CXR dataset, and 100 normal and 200 abnormal cases from the IND1 dataset. We ensure coverage of multiple abnormal cases for both datasets, because we found classification quality to vary significantly across conditions. It is also worth noting that the same set of cases was annotated in both the error correction and pairwise preference tasks. For the MIMIC-CXR dataset, we include cases annotated in the human evaluation of the previous work²² that survived the filtering stage described below.

Annotation interface. We use an internal platform for data collection to perform our expert evaluation. Extended Data Fig. 1 illustrates the labeling interfaces used by our raters to perform the pairwise preference and error correction tasks. The annotators were provided with the following descriptions of the respective tasks along with screenshots of examples:

(1) Instructions for Pairwise Preference Task

You are provided with:

- The CXR image
- Two radiology reports for this image, each consisting of the findings and impression sections.

Your task is to help us assess the relative usefulness of the radiology reports. An example with detailed instructions is shown in Extended Data Fig. 1a.

(2) Instructions for Error Correction Task

You are provided with:

- The CXR image
- A radiology report for this image, consisting of the findings and impression sections.

Your task is to help us assess the accuracy of the radiology report in detail. You will be asked if there is any part of the report that you do not agree with and, if so, you will then be asked to (a) select the passage that they disagree with, (b) select the reason for disagreement ('finding I do not agree with is present'; 'incorrect location of finding'; 'incorrect severity of finding'), (c) specify whether the error is clinically significant or not, and (d) provide a replacement for the selected passage. An error should be labeled as clinically significant if it is potentially harmful and could change treatment/outcome for a patient. An example is shown in Extended Data Fig. 1b.

In addition, we addressed their questions on an as-needed basis through emails.

All data were stored in the Digital Imaging and Communications in Medicine (DICOM) format and deidentified before transfer to the external radiologists for annotation. Experts were asked to confirm whether the image provided to them for each task was of sufficient quality for them to complete the task. In three MIMIC-CXR cases, one of the four raters nominated not to complete the task. In those instances, the entire case was discarded. After these exclusions, the MIMIC-CXR evaluation set consisted of 32 normal cases and 271 abnormal cases, with abnormal conditions occurring at the following frequencies (in parentheses): lung opacity (132), cardiomegaly (123), support devices (134), pleural effusion (100), atelectasis (95), edema (75), enlarged cardiomeastinum (68), pneumonia (46), consolidation (25), lung lesion (17), pneumothorax (13), fracture (10), pleural other (8), with many abnormal cases containing more than one condition. Evaluators were given full resolution X-ray images but were not given Indication data or clinical history data, or any other data about the possible origin of a report, consistent with the model task formulation in our study and with previous studies^{10,15,29,40}.

Pairwise preference test. Clinicians were then asked, 'If you had to choose one of these two reports to go into the Picture Archiving and Communication System (PACS) system and be used downstream for

the care of this patient, which would be best for the patient?'. For each case, the raters are unaware of which report is the original and they are not aware that one of the reports was generated by our AI system. We note that the assignment of the original and the generated reports to option A and B is completely random for each case.

Error correction. Before each annotation task, clinicians are asked whether the presented image is of sufficient quality for them to complete the task. They are then asked whether there is any part of the report that they do not agree with and, if so, are asked to (1) select the passage that they disagree with, (2) select the reason for disagreement ('finding I do not agree with is present'; 'incorrect location of finding'; 'incorrect severity of finding'), (3) specify whether the error is clinically significant or not, and (4) provide a replacement for the selected passage. We instruct the raters beforehand that a clinically significant error is one that is potentially harmful or influences the downstream clinical decision (for example, treatment) for the patient. We note that the raters evaluate both the GT reports written by an expert and the ones generated by our model, but without the knowledge of their sources. As the raters performing this task are different from the ones that wrote the original reports, this would also allow us to measure the degree of human errors in report writing. Importantly, our evaluation differs from the previous work²² where the original report was additionally provided as a reference and, hence, was assumed to be accurate.

Clinician–AI collaboration. We use the pairwise preference interface described above and we ensure that the clinician that produces a clinician–AI report is excluded from the group that performs the preference test for that report. We exclude reports where the raters did not provide replacement sentences as instructed in the error correction task (seven MIMIC-CXR instances and four IND1 instances). We evaluate expert preferences for the IND1 and MIMIC-CXR datasets, and for each report, we collect preferences from four radiologists (two from our India cohort and two from our US cohort). Identical to the previous setup, the raters do not know which report corresponds to the original GT and which was initially generated by the AI model. In all these cases, we report rater preferences for reports that were subject to editing by clinicians, so that comparisons shed light on the effect of clinician–AI collaboration.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

'MIMIC-CXR'²⁷, one of the real-world datasets used in the development of Flamingo-CXR is accessible by researchers and can be downloaded from <https://physionet.org/content/mimic-cxr> upon completion of the required training. IND1, the other deidentified chest X-ray dataset used in this study cannot be made publicly available because the authors do not have the rights to do so. Interested researchers should contact info@apollohospitals.com to inquire about access to the IND1 dataset; requests will be subject to Apollo's consideration and applicable ethical and legal requirements. The radiologist ratings and generated reports are not publicly available because these are inextricably linked to the IND1 dataset and MIMIC-CXR dataset as described in the Methods. Further inquiries about our benchmarking procedures and data analysis may be addressed to the corresponding authors with a maximum response time of two weeks.

Code availability

For reproducibility, we have documented the technical details of the implementation while keeping the paper accessible to a clinical and general scientific audience. Several major components of our work are available in open source repositories, such as the Haiku library

(<https://github.com/google-deepmind/dm-haiku>). Our work builds upon Flamingo, for which implementational details have been described extensively in the corresponding publication⁸ and an open source implementation of the base model; for instance, the Open-Flamingo project available at https://github.com/mlfoundations/open_flamingo. Other components used in our work cannot be shared publicly because of their proprietary nature.

References

40. Yan, A. et al. Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* 4009–4015 (2021).
41. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs & outputs. In *International Conference on Learning Representations* (ICLR, 2022).
42. Irvin, J. et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 33 590–597 (2019).
43. Loshchilov, I. & Hutter, F. Fixing weight decay regularization in Adam. Preprint at <https://arxiv.org/abs/1711.05101v2> (2018).
44. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. Preprint at <https://arxiv.org/abs/1904.09751> (2019).
45. Ahn, J. S. et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. *JAMA Netw. Open* **5**, e2229289 (2022).
46. Ramesh, V., Chi, N.A. & Rajpurkar, P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. *Proc. Mach. Learn. Res.* **193**, 456–473 (2022).
47. Vedantam, R., Zitnick, C. L. & Parikh, D. CIDEr: consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4566–4575 (2015).
48. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (Association for Computational Linguistics, 2002).
49. Lin, C.-Y. in *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
50. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. T. On faithfulness and factuality in abstractive summarization. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL, 2020)*.
51. Pătrăucean, V. et al. Perception Test: a diagnostic benchmark for multimodal video models. *Adv. Neural Inform. Proc. Syst.* **36** (2024).
52. Horvitz, D. G. & Thompson, D. J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **47**, 663–685 (1952).

Acknowledgements

This research was funded by Google DeepMind and Google Research. We would like to thank many colleagues for useful discussions, suggestions, feedback and advice, including C. Kelly, G. S. Corrado, J.

Krause, N. Tomasev, O. Vinyals, S. Mohamed, R. Hadsel, Z. Ghahramani, T. Cemgil and Y. Chen.

Author contributions

A.K., P.K. and P.S.-H. initiated the project. R.T., D.G.T.B., I.K., A.K., A. Sellergren, S.G., S.D., A. See, D.B., P.-S.H. and J.W. contributed to the design of the method and experiments. S.G., S.D., A. See, D.G.T.B., P.S.-H., J.W., A.K. and V.N. contributed to the initial problem formulation and engineering setup. R.T., D.G.T.B. and I.K. revised the problem formulation and the engineering setup. I.K., R.T., D.G.T.B., A. Sellergren, S.G., S.D., A. See, P.S.-H. and J.W. contributed to software engineering. R.T. and I.K. evaluated and developed the final report generation model. S.G. ingested and prepared the MIMIC-CXR dataset. S.R.K. provided support for use of the IND1 dataset. A.K., C.L. and S.D. designed the human evaluation rubric. A. Sellergren, I.K. and R.T. implemented the labeling pipeline for human evaluation. S. Man, R.L., D.G.T.B. and R.T. coordinated the labeling jobs. I.K., D.G.T.B. and R.T. contributed to the evaluation of the work and performed analysis. D.G.T.B., I.K., A.K., S.D. and R.T. contributed to the interpretation of the results. T.T., S.A., M.S., V.N., D.B., K.S., Z.A., P.K., Y.M., J.B., Y.L., S.M.A.E. and S.S. advised on the work. R.T., D.G.T.B., I.K. and A.K. wrote the paper. A. Sellergren, S.D., J.W., T.T., V.N., S. Mahdavi, R.M., S.A., S.M.A.E. and A.K. revised the manuscript. D.G.T.B., I.K. and R.T. incorporated feedback during the review process.

Competing interests

This study was funded by Google LLC and/or a subsidiary thereof ('Google'). R.T., D.G.T.B., A. Sellergren, S.G., S.D., A. See, J.W., C.L., T.T., S.A., M.S., R.M., R.L., S. Man, Z.A., S. Mahdavi, Y.M., J.B., S.M.A.E., Y.L., S.S., V.N., P.K., P.S.-H., A.K. and I.K. are employees of Google and may own stock as part of the standard compensation package. D.B. was a Google employee and is currently an employee of GlaxoSmithKlein AI division and may own stock as part of the standard compensation package. Similarly, K.S. was a Google employee and may own stock, but is currently an employee of OpenAI.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41591-024-03302-1>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-024-03302-1>.

Correspondence and requests for materials should be addressed to Ryutaro Tanno, David G. T. Barrett, Alan Karthikesalingam or Ira Ktena.

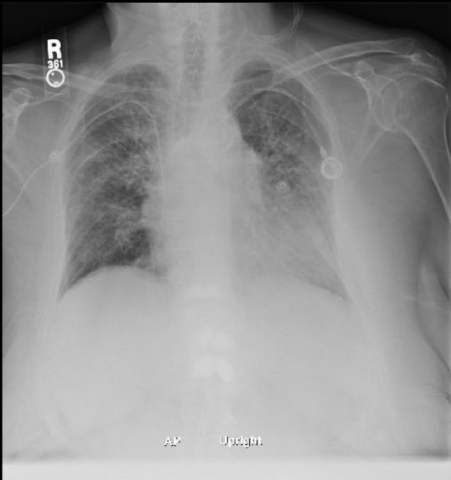
Peer review information *Nature Medicine* thanks Jarrel Seah and Fredrik Strand for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

a. Labelling interface for pairwise preference test

Report A: FINDINGS: A Port-A-Cath terminates at the cavoatrial junction. The cardiac, mediastinal and hilar contours appear stable. There is no pleural effusion or pneumothorax. A mild interstitial abnormality suggests pulmonary vascular congestion, but there is no focal opacification. IMPRESSION: Findings suggesting mild vascular congestion.

Report B: FINDINGS: AP upright and lateral chest radiographs were obtained. Known interstitial lung disease contributes to a bilateral perihilar interstitial abnormality. In addition to the chronic findings there is bilateral ground-glass opacity and interstitial thickening, predominantly radiating from the hila. Cardiomegaly remains moderate. Aortic arch calcifications are unchanged. A right-sided PICC line terminates in the low SVC. A left chest Port-A-Cath terminates in the right atrium. Vertebroplasty changes are stable. IMPRESSION: New pulmonary parenchymal abnormalities on top of chronic pulmonary fibrosis most likely represents pulmonary edema. Infection is less likely.



*** Required**

If you had to choose one of these two reports to go into the PACS system and be used downstream for the care of this patient, which would be best for the patient? *

☐ Report A

☒ Report B

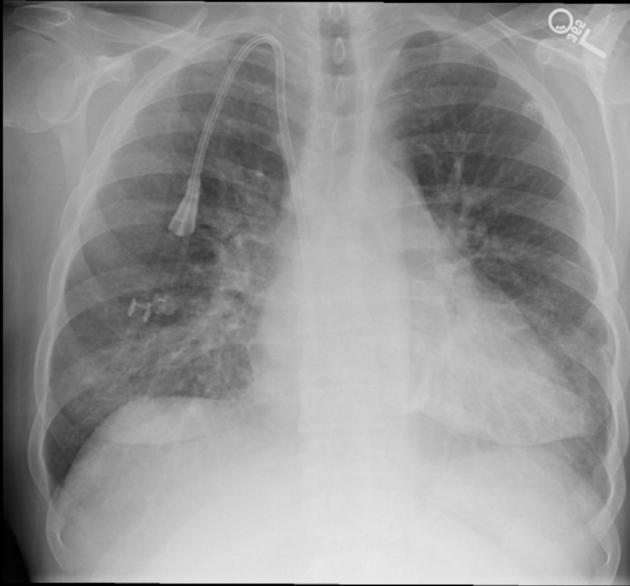
☐ Neither is better than the other

Please explain in 1-3 sentences why you have selected the above option. *

In report B there is mention of position of PICC line and report is more accurate in mentioning pulmonary edema changes over chronic pulmonary fibrosis.

b. Labelling interface for error correction

[1] FINDINGS: PA and lateral views of the chest provided. [2] Right IJ access dialysis catheter again noted with tip in the low SVC. [3] Cardiomegaly is again noted with hilar congestion and moderate pulmonary edema. [4] No large effusion or pneumothorax. [5] No convincing signs of pneumonia. [6] Mediastinal contour is stable. [7] Bony structures are intact. [8] IMPRESSION: Moderate pulmonary edema.



Zoom: 26.03%

*** Required**

Is the image quality sufficient to perform this task fully? *

Is there any part of the report that you do not agree with? *

For each disagreement, select the passage you disagree with.

Disagreement passage #1

For each disagreement, enter the number of the sentence you disagree with.

Why do you disagree with this passage? *

☐ Finding I do not agree is present (a)

☒ Incorrect location of finding (l)

☐ Incorrect severity of finding (s)

This error is: *

☐ clinically significant (s)

☒ clinically insignificant (l)

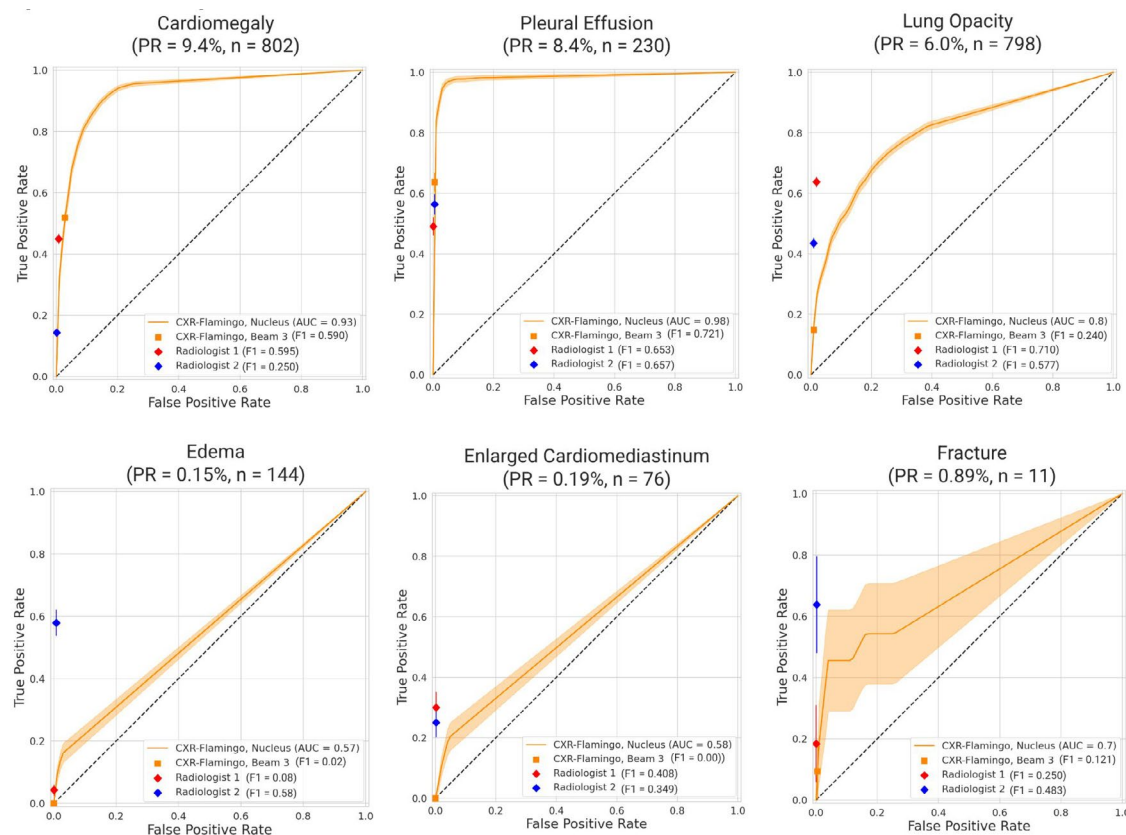
Write what you would put in place of the selected passage *

PA view of the chest provided.

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Labelling interface. (a) In the labelling interface for the pairwise preference test, raters are provided with (i) a frontal view (PA or AP) in the original resolution, (ii) a radiology report generated by our AI system and (iii) the original report written by a radiologist, and are asked to provide their preference. For each case, the raters are unaware of which report is the ground-truth and which one is generated by our model, and are requested to describe their preference out of three options; report A, report B, or equivalence between the two (that is, 'neither is better than the other'). The interface allows the raters to zoom in and out on the image as needed. They are additionally asked to provide an explanation for their choice. (b) In the labelling interface for the error correction task, raters are provided with (i) the chest X-ray image (a frontal

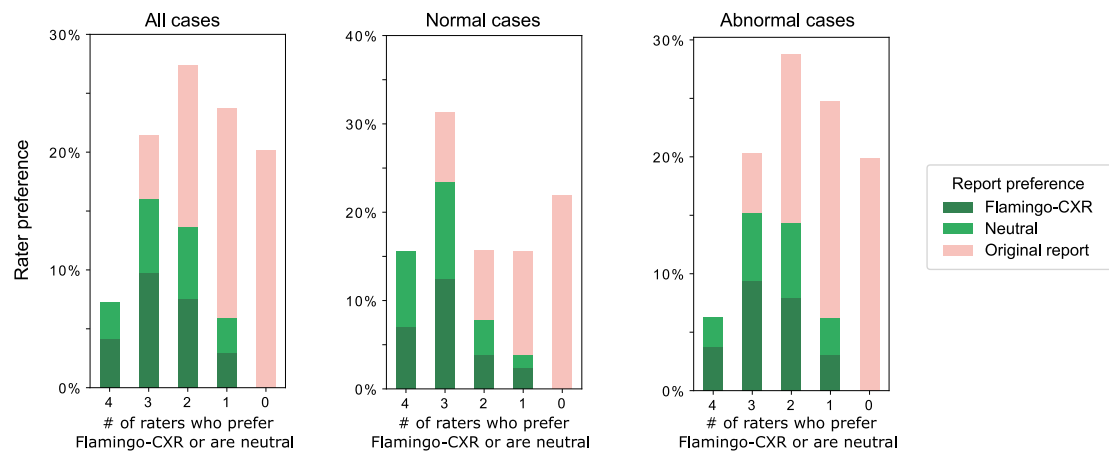
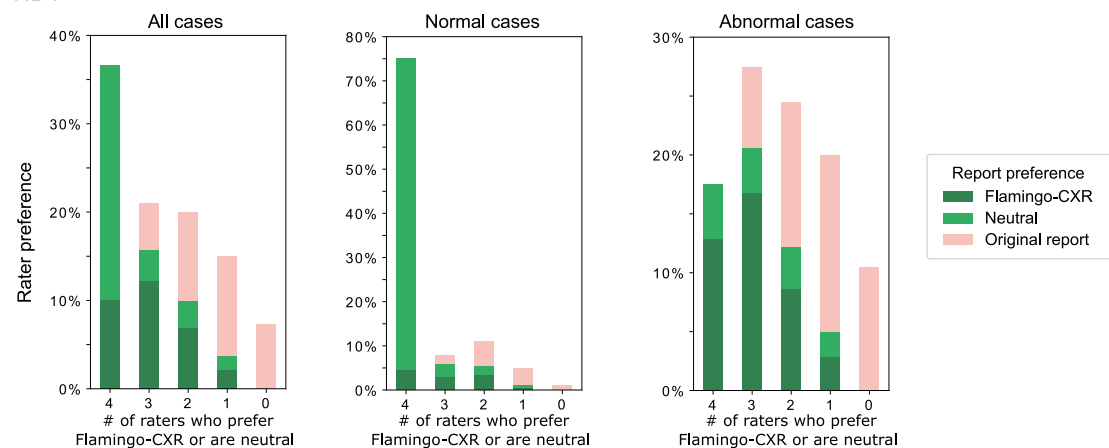
view) and (ii) a radiology report for this image, consisting of the findings and impression sections. Their task is to assess the accuracy of the given radiology report by identifying errors in the report and correcting them. Before each annotation task, clinicians are asked whether the presented image is of sufficient quality for them to complete the task. They are then asked whether there is any part of the report that they do not agree with and, if so, are asked to (a) select the passage that they disagree with, (b) select the reason for disagreement (finding I do not agree with is present; incorrect location of finding; incorrect severity of finding), (c) specify whether the error is clinically significant or not, and (d) provide a replacement for the selected passage.



Extended Data Fig. 2 | Detection accuracy per condition on the IND1 dataset.

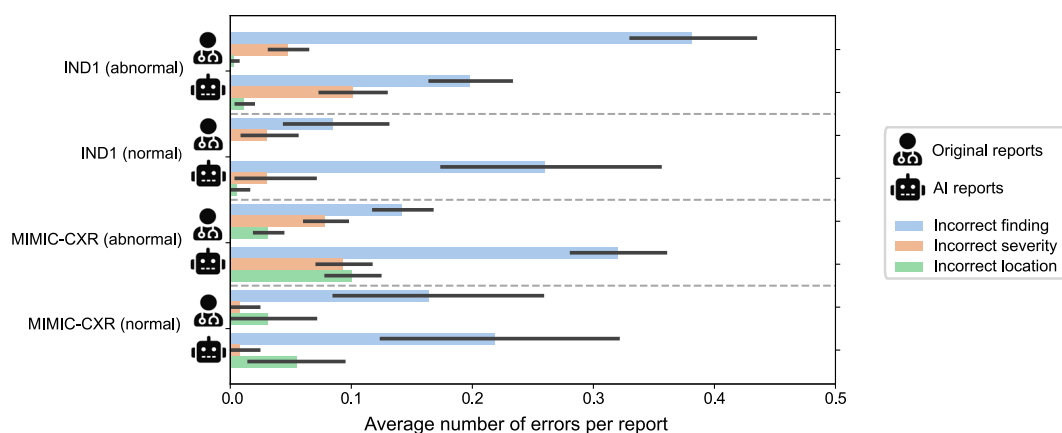
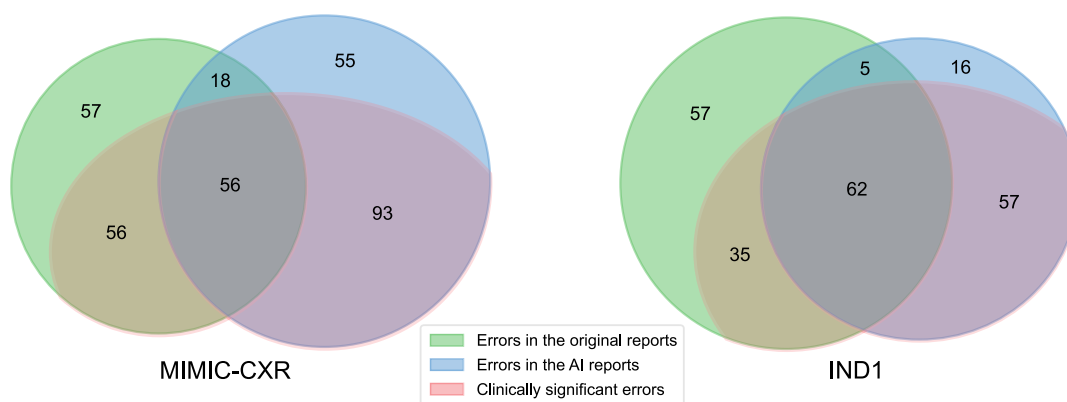
The receiver operating characteristic (ROC) curve of the Flamingo-CXR report generation model, shown along with the true positive rate (TPR) and false positive rate (FPR) pairs for two certified radiologists are shown for 6 conditions

for which the expert labels were collected. The operating point of our model with the default inference scheme (Beam 3) is also shown. Error bars represent 95% confidence intervals (calculated using bootstrapping with 1000 repetitions).

a. MIMIC-CXR**b. IND1**

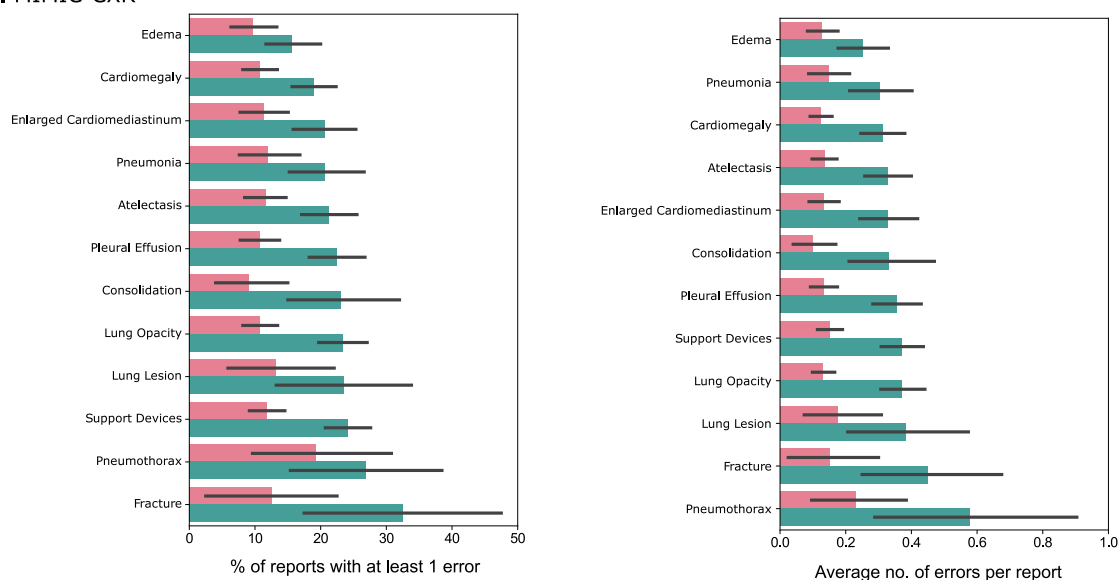
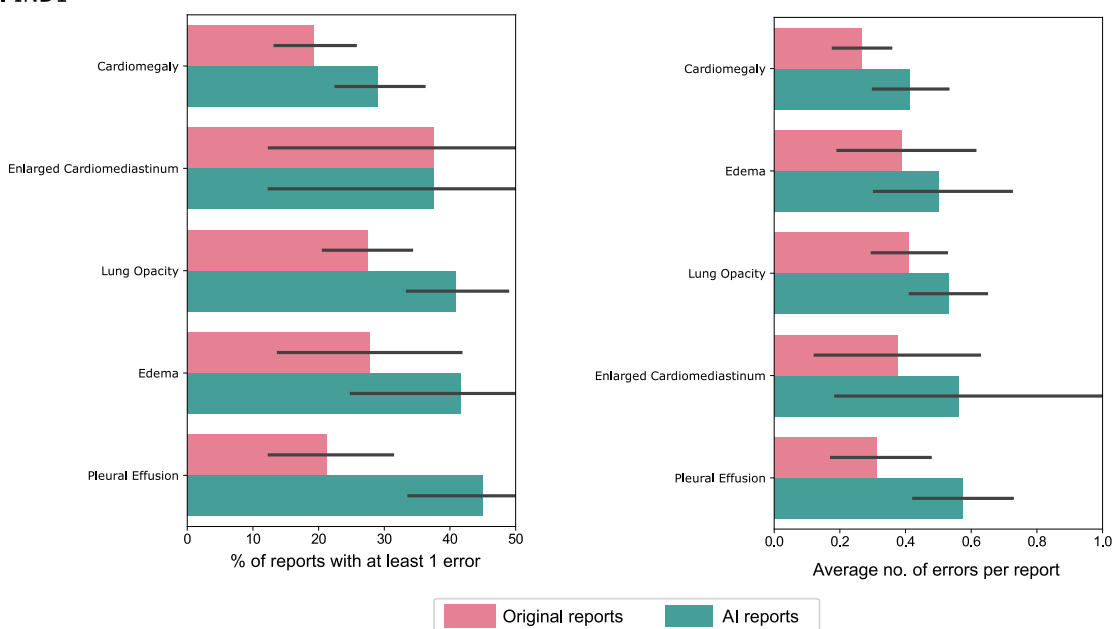
Extended Data Fig. 3 | Subgroup analysis of preferences for MIMIC-CXR and IND1. Here the expert preference data presented in Fig. 3 is analysed further, with preferences shown separately for Flamingo-CXR reports, ground truth reports and neutral preference between reports, for (a) MIMIC-CXR reports and (b) IND1

reports. As before, reports are grouped according to the level of agreement between reviewers who rate Flamingo-CXR reports as equivalent or better than ground truth reports. Preferences are further grouped into normal and abnormal subsets.

a. Types of errors found in the original reports and the AI-generated reports**b. Intersection of errors between the original reports and the AI-generated reports**

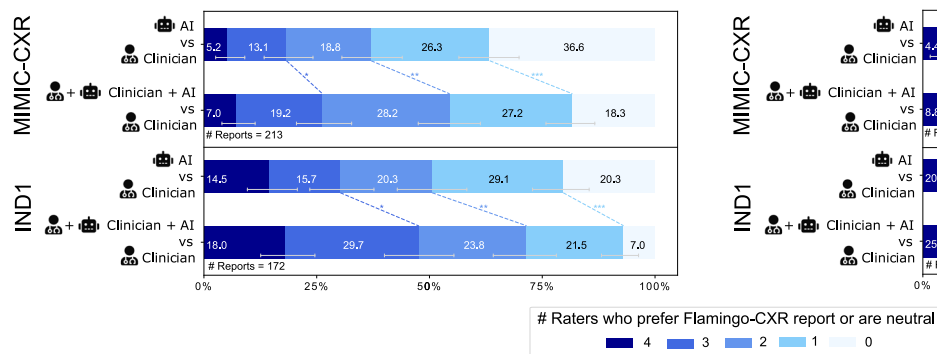
Extended Data Fig. 4 | Types of errors found in the original reports and the AI-generated reports. (a) During the error correction evaluation, we ask expert raters to explain the identified issues in reports based on the following taxonomy: (i) incorrect findings, (ii) incorrect severity (for example, mild vs. severe pulmonary edema), (iii) incorrect location of finding (for example, left- vs. right-sided pleural effusion). The figure shows the distributions of these error types for the normal and abnormal cases separately in the IND1 and MIMIC-CXR datasets. Data is presented as mean values and 95% confidence intervals across

cases are also shown. In total, there are 34 normal and 272 abnormal cases from the MIMIC-CXR dataset, and 100 normal and 200 abnormal cases from the IND1 dataset. (b) Venn diagrams of error counts for reports that contain at least one error, for the MIMIC-CXR dataset and the IND1 dataset. The intersection between the blue and the green segments indicates the number of cases where both the AI-generated report and the ground truth contained errors. The red segment indicates the cases where at least one clinically significant error is detected.

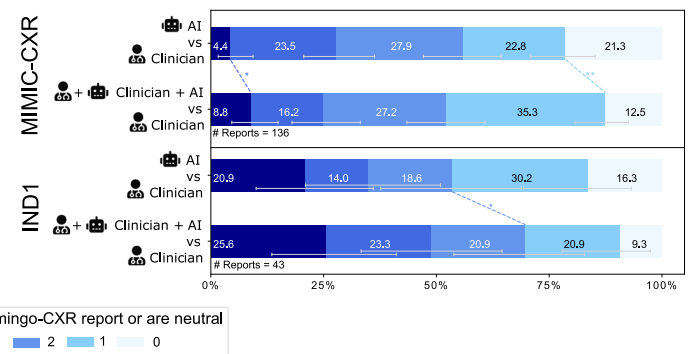
a. MIMIC-CXR**b. IND1**

Extended Data Fig. 5 | Average number of clinically significant errors and percentage of reports with at least one error reported by experts in human-written and AI-generated reports across conditions for the MIMIC-CXR and IND1 datasets. (a) For MIMIC-CXR, the average number of clinically significant errors in reports that are capturing cases with *pneumothorax* is almost double the number of those with *edema*, but for most other conditions the occurrence of errors does not vary significantly. It is worth noting that the condition labels for MIMIC-CXR cases are obtained using CheXpert³² on the original human-written reports. Additionally, if more than one condition is associated with a particular

chest X-ray image (which is often the case), the clinically significant errors on the corresponding reports are reported for all of these conditions. **(b)** For IND1, we do not observe striking differences across conditions in terms of clinically significant errors reported in the AI-generated reports, even though there are more errors on average reported for cases with *pleural effusion* than those with *cardiomegaly*. Interestingly, no errors are reported in cases with *fracture*, so we omit this condition from the figure. These findings indicate that condition prevalence in the training data does not necessarily affect report quality.

a. With clinically significant errors

Extended Data Fig. 6 | Clinician-AI collaboration and clinically significant errors. Subgroup analysis of the data presented in Fig. 5 illustrates that (a) clinician-AI collaboration produced an improvement in ratings for the subgroup of AI reports that had clinically significant errors (with MIMIC-CXR p values given by $p^* = 2.6 \times 10^{-3}$, $p^{**} = 1.5 \times 10^{-7}$, $p^{***} = 2.9 \times 10^{-8}$ and with IND1 p values given by $p^* = 6.3 \times 10^{-7}$, $p^{**} = 4.0 \times 10^{-8}$, $p^{***} = 1.3 \times 10^{-5}$), whereas (b), there was little or no improvement for the subgroup of AI reports that did not have clinically significant errors (with MIMIC-CXR p values given by $p^* = 1.2 \times 10^{-2}$, $p^{**} = 1.2 \times 10^{-2}$

b. Without clinically significant errors

and with IND1 p values given by $p^* = 3.2 \times 10^{-2}$). As before, significant differences ($p < 0.05$) between clinician-AI results and AI-only results calculated using a one-sided Chi-squared are indicated by asterisks. This suggests that the positive impact of clinician-AI collaboration is largely attributable to edits in AI reports that had clinically significant errors. Data for all panels is presented as mean values and error bars show 95% confidence intervals for the cumulative preference scores.

Extended Data Table 1 | Examples of reports that contain references to prior measurements both in the original clinician reports and in the predicted reports from Flamingo-CXR trained with/without pre-processing

Original	Prediction (trained with pre-processing)	Prediction (trained without pre-processing)
<p>FINDINGS: AP view of the chest. The patient has been intubated since prior exam. Right PICC is seen with tip at the upper SVC. Relatively low lung volumes are seen. The lungs however remain clear without consolidation, effusion or pulmonary vascular congestion . Cardiac silhouette appears moderately enlarged, likely accentuated due to low lung volumes and AP technique.</p>	<p>FINDINGS: Single AP upright portable view of the chest was obtained. A right-sided PICC is seen, terminating at the cavoatrial junction/right atrium. The cardiac silhouette appears mildly enlarged. Mediastinal contours are stable. There is no overt pulmonary edema. No large pleural effusion is seen, although a trace right pleural effusion would be difficult to exclude.</p>	<p>FINDINGS: As compared to the previous radiograph, the patient has been intubated. The tip of the endotracheal tube projects approximately 3.5 cm above the carina. There is no evidence of complications, notably no pneumothorax. The other monitoring and support devices are constant, with the exception of the nasogastric tube that has been newly inserted. The course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. Unchanged evidence of bilateral parenchymal opacities and of bilateral pleural effusions. Unchanged borderline size of the cardiac silhouette.</p>

The descriptions that mention prior data are highlighted.

Extended Data Table 2 | Automated report generation metrics

a. Automated report generation metrics on MIMIC-CXR

Model	Sections	NLG Metrics			Clinical Metrics		
		CIDEr	BLEU4	Rouge	F1 (all)	F1 (top 5)	Radgraph
CXR-RePaiR (Endo et al., 2021)	F	-	0.021	0.143	0.281	-	0.091
M^2 Transformer (Miura et al., 2021)	F	0.509	0.114	-	-	0.567	0.220
RGRG (Tanida et al., 2023)	F	0.495	0.126	0.264	0.447	0.547	-
METransformer (Wang et al., 2023a)	F	0.362	0.124	0.291	0.311	-	-
Med-PaLM-M, 12B (Tu et al., 2023)	F	0.234	0.104	0.262	0.514	0.565	0.252
R2Gen (Chen et al., 2020)	F + I	-	0.103	0.277	0.228	0.346	0.134
WCT (Yan et al., 2021)	F + I	-	0.144	0.274	0.294	-	0.143
CvT-21DistillGPT2 (Nicolson et al., 2023)	F + I	0.361	0.124	0.285	0.384	-	0.154
BioVil-T (Bannur et al., 2023)	F + I	-	0.092	0.296	0.317	-	-
R2GenGPT (Wang et al., 2023b)	F + I	0.269	0.134	0.297	0.389	-	-
Flamingo-CXR (Ours)	F + I	0.138	0.101	0.297	0.519	0.580	0.205

b. Automated report generation metrics on IND1

Model	NLG Metrics			Clinical Metrics		
	CIDEr	BLEU4	Rouge-L	F1 (all)	F1 (top 3)	Radgraph
Flamingo-CXR (Ours)	5.158	0.724	0.851	0.463	0.512	0.805
Radiologist 1	-	-	-	0.621	0.657	-
Radiologist 2	-	-	-	0.467	0.476	-

(a) Comparison of automatic report generation metrics on the MIMIC-CXR dataset. The column ‘Sections’ indicates which sections of the radiology reports are generated by the respective models; ‘F’ indicates FINDINGS and ‘I’ indicates IMPRESSIONS sections. Note that the metrics are retrieved from the corresponding publications. For all metrics, the higher (the bluer) the better, and the best results are shown in bold. (b) Automated report generation metrics on the IND1 dataset. We note that there are no published report generation metrics due to the private nature of the dataset. The disease classification accuracy (F_1 scores) are also computed for two radiologists.

Extended Data Table 3 | Examples of cases with clinically significant errors found in the ground truth radiology report, but not in the predicted Flamingo-CXR report

Predicted report	Radiologist report	Replacement	Disagreement reason
<p>FINDINGS: Single supine AP portable view of the chest was obtained. No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. The cardiac silhouette is top normal. The aorta is calcified and tortuous. No displaced fracture is seen.</p> <p>IMPRESSION: No acute cardiopulmonary process.</p>	<p>FINDINGS: The lungs are low in volume but otherwise clear. Left hemidiaphragm is somewhat obscured in its lateral-most component, though this could be projectional. The left lung base is poorly imaged. There is no definite pleural effusion or pneumothorax. Stable marked cardiomegaly is noted.</p> <p>IMPRESSION: Questionable opacity in left base. When the patient's clinical status improves, repeat evaluation by PA and lateral chest radiograph is recommended to exclude a pleural effusion or left basilar parenchymal process.</p>	Increased cardiac silhouette is likely due to position and technical region and not true cardiomegaly.	Finding I do not agree is present.
<p>FINDINGS: Frontal and lateral radiographs of the chest demonstrate stable post-radiation paramediastinal fibrosis and scarring in the right upper lobe. There is a small right-sided pleural effusion with adjacent atelectasis. The cardiomeastinal and hilar contours are unchanged. There is no pneumothorax.</p> <p>IMPRESSION: Small right-sided pleural effusion with adjacent atelectasis.</p>	<p>FINDINGS: An extensive right hilar lung mass is associated with radiation fibrosis, better delineated on CT _____. An additional component of postobstructive pneumonia may be present. Retrocardiac opacity, left pleural effusion, and left pleural thickening are also new. No pneumothorax is present.</p> <p>IMPRESSION: 1. Large right hilar lung mass and radiation fibrosis. Additional post-obstructive pneumonia in the right upper and lower lobes is possible but hard to delineate. 2. New left retrocardiac opacity, small left effusion, and pleural thickening. Findings were discussed with ___, RN, via telephone at ___ and again with Dr ___ at ___.</p>	In addition, right pleural effusion versus thickening.	Incorrect severity of finding.
<p>FINDINGS: Severe cardiomegaly is redemonstrated. The mediastinal and hilar contours are unchanged. There is mild pulmonary edema, worse in the interval. No focal consolidation, pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.</p> <p>IMPRESSION: Severe cardiomegaly with mild pulmonary edema, worse in the interval.</p>	<p>FINDINGS: The lungs are well expanded and clear. Area of increase density overlying the right hilum with a sharp lower margin is of unclear clinical significance. Severe cardiomegaly is reidentified. The hilar contours are unremarkable. There is no pleural effusion or pneumothorax.</p> <p>IMPRESSION: 1. Area of increase density overlying the right hilum with a sharp lower margin is of unclear clinical significance. Chest CT is recommended for further assessment. 2. Severe cardiomegaly, unchanged. The impression was entered as an urgently flagged wet read on the ED dashboard by Dr ___ on ___ at 9:05 am after discussion with the attending as the patient was still in the ED.</p>	I would have written - mild pulmonary vascular congestion.	Finding I do not agree is present.

Clinically significant errors in the ground truth radiology report are highlighted in red. Information that was identified to be missing in the ground truth report but is represented in the Flamingo-CXR report is highlighted in green.

Extended Data Table 4 | Examples of cases with clinically significant errors found in the Flamingo-CXR report, but not in the original ground-truth report

Predicted report	Radiologist report	Replacement	Disagreement reason
FINDINGS: Dual lead left-sided pacemaker is seen with leads extending the expected positions of the right atrium and right ventricle. No focal consolidation is seen. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable. IMPRESSION: No acute cardiopulmonary process.	FINDINGS: Single lead pacemaker in situ with the lead tip in the right ventricle. No cardiomegaly. No features of cardiac decompensation. Prominent pulmonary arteries suggesting pulmonary arterial hypertension. No pleural effusion. Consolidation in the left lower lobe. IMPRESSION: Suspected left lower lobe pneumonia.	Single chamber pace maker with tip is in right ventricle.	Finding I do not agree is present.
FINDINGS: Lung volumes are low. Heart size is accentuated as a result, and appears mild to moderately enlarged. The aorta is tortuous and diffusely calcified. There is crowding of the bronchovascular structures with mild pulmonary edema noted. No large pleural effusion or pneumothorax is seen. Patchy opacities in the lung bases likely reflect areas of atelectasis. IMPRESSION: Mild pulmonary edema. Low lung volumes.	FINDINGS: Persistent pulmonary opacities, vascular engorgement and septal lines reflect mild pulmonary edema. Small left pleural effusion cannot be excluded. Low lung volumes limit assessment of cardiomeastinal silhouette though the cardiac size appears mildly enlarged. IMPRESSION: Unchanged mild pulmonary edema with likely small left pleural effusion.	Only aortic knuckle calcification is present.	Incorrect severity of finding.
FINDINGS: Frontal and lateral chest radiographs were obtained. A right-sided Port-A-Cath terminates in the lower SVC. The lungs are fully expanded and clear. The cardiomeastinal silhouette, hilar contours, and pleural surfaces are normal. There is no pleural effusion or pneumothorax. IMPRESSION: No focal consolidation to suggest pneumonia.	FINDINGS: A right-sided Port-A-Cath tip sits in the lower portion of the SVC. The heart and mediastinal contours are within normal limits. The lungs are largely clear with only minimal atelectasis in the right base in accordance with a small right pleural effusion. There is no pneumothorax. IMPRESSION: Small right pleural effusion with associated atelectasis; no pneumothorax.	There is mild right sided pleural effusion.	Finding I do not agree is present.
FINDINGS: The lungs are well expanded and clear. The hila and pulmonary vasculature are normal. No pleural effusions or pneumothorax. The cardiomeastinal silhouette is normal. A left pectoral pacemaker is seen with transvenous leads in the right atrium and right ventricle. IMPRESSION: No acute cardiopulmonary process.	FINDINGS: The lungs appear clear. A pacemaker is seen projecting over the left chest with a wire appropriately placed in the right atrium. The cardiomeastinal silhouette, hilar contours, and pleural structures are normal. No pneumothorax or pleural effusion. Other than the pacemaker, no radio-opaque metallic foreign object is identified in chest radiograph. IMPRESSION: 1. Pacemaker seen projecting over the left chest with a wire appropriately placed in the right atrium. Other than the pacemaker, no radiopaque metallic foreign object is identified. 2. No acute cardiopulmonary process.	Single chamber pace maker with lead in right atrium.	Incorrect location of finding.

Clinically significant errors in the Flamingo-CXR report are highlighted in red. Information that was identified to be missing in the Flamingo-CXR report but is represented in the ground truth radiologist report is highlighted in green.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of all covariates tested
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	We implemented the annotation tool for conducting the expert evaluation using an internal software.
Data analysis	We used Python and standard visualisation/analysis toolkits (e.g., matplotlib and scipy) for data analysis. We also used CheXPert labeller, an open-sourced NLP software for pre-processing radiology reports.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Below is the data availability statement we included in our manuscript:
"MIMIC-CXR27, one of the real-world datasets used in the development of Flamingo-CXR is accessible by researchers and can be downloaded from <https://physionet.org/content/mimic-cxr> upon completion of the required training. IND1, the other de-identified chest X-ray dataset used in this study cannot be made

publicly available because the authors do not have the rights to do so. Interested researchers should contact info@apollohospitals.com to inquire about access to the IND1 dataset; requests will be subject to Apollo's consideration and applicable ethical and legal requirements. Further enquiries about the data used in this study may be addressed to the corresponding authors with a maximum response time of two weeks. "

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	This information has not been collected.
Population characteristics	We evaluate our AI system with a group of radiologists located in two different countries, namely India and the US.
Recruitment	We ensured that the recruited radiologists are capable for performing the evaluation task reliably through a series of written/verbal trainings and by making sure that they hold board certifications of the countries of their residence. We recruited experts from India and the USA to ensure diversity of raters.
Ethics oversight	The use of de-identified retrospective datasets was reviewed by Advarra IRB (Columbia, MD), which determined that it was exempt from further review under 45 CFR 46. The involvement of clinicians in this study, using the same de-identified retrospective data is also covered in this waiver.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<p>Sample size was determined by the maximum number of reports and replications viable a budget constraint, while making sure that the sample size (550 reports x 4 annotations per example) in this study exceeds the relevant prior publications such as [1] where 500 reports with 3 annotations per report are used, and [2] where 250 reports x 1 annotation per report are used:</p> <p>[1] J. Huang, L. Neill, M. Wittbrodt, D. Melnick, M. Klug, M. Thompson, J. Bailitz, T. Loftus, S. Malik, A. Phull, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Network Open, 6(10):e2336100–e2336100, 2023.</p> <p>[2] T. Tu, S. Azizi, D. Driess, M. Schaeckermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical AI. NEJM AI, 1(3):Aloa2300138, 2024a.</p>
Data exclusions	There were a several cases that we excluded from the clinician-AI rater study. Concretely, there were 4 IND1 cases and 7 MIMIC-CXR cases where the radiologists did not complete part (d) of the edit correction task in accordance with the instructions; and these were excluded from the clinician-AI rater study.
Replication	Each evaluation task is conducted by 4 experts (2 based in the US and 2 in India).
Randomization	Both examples / human participants were randomly allocated to the annotation tasks.
Blinding	The investigators were completely blinded to the group allocation

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging