MAIRA-2: Grounded Radiology Report Generation

Shruthi Bannur[*1], Kenza Bouzid[*1], Daniel C. Castro[1], Anton Schwaighofer[1], Sam Bond-Taylor[1], Maximilian Ilse[1], Fernando Pérez-García[1], Valentina Salvatelli[1], Harshita Sharma[1], Felix Meissen[1], Mercy Ranjit[2], Shaury Srivastav[2], Julia Gong[3], Fabian Falck[1], Ozan Oktay[1], Anja Thieme[1], Matthew P. Lungren[4], Maria Teodora Wetscherek[1,5], Javier Alvarez-Valle[°1], and Stephanie L. Hyland[°1]

[1]Microsoft Research Health Futures
[2]Microsoft Research India
[3]Microsoft Azure AI
[4]Microsoft Health and Life Sciences
[5]Department of Radiology, Addenbrooke's Hospital, Cambridge University Hospitals

Abstract

Radiology reporting is a complex task that requires detailed image understanding, integration of multiple inputs, including comparison with prior imaging, and precise language generation. This makes it ideal for the development and use of generative multimodal models. Here, we extend report generation to include the localisation of individual findings on the image – a task we call grounded report generation. Prior work indicates that grounding is important for clarifying image understanding and interpreting AI-generated text. Therefore, grounded reporting stands to improve the utility and transparency of automated report drafting.

To enable evaluation of grounded reporting, we propose a novel evaluation framework – RadFact– leveraging the reasoning capabilities of large language models (LLMs). RadFact assesses the factuality of individual generated sentences, as well as correctness of generated spatial localisations when present.

We introduce MAIRA-2, a large multimodal model combining a radiology-specific image encoder with a LLM, and trained for the new task of grounded report generation on chest X-rays. MAIRA-2 uses more comprehensive inputs than explored previously: the current frontal image, the current lateral image, the prior frontal image and prior report, as well as the *Indication*, *Technique* and *Comparison* sections of the current report. We demonstrate that these additions significantly improve report quality and reduce hallucinations, establishing a new state of the art on findings generation (without grounding) on MIMIC-CXR while demonstrating the feasibility of grounded reporting as a novel and richer task.

1 Introduction

Medical imaging is central to the safe and effective delivery of modern medicine (UK HSA, 2022). It is integral to numerous treatment pathways, providing the necessary insights for precise diagnoses and therapeutic decisions. Nonetheless, the escalating demand for imaging services is surpassing the capacity of radiologists to maintain a high level of proficiency in image reporting (Fischetti et al., 2022; Kalidindi & Gandhi, 2023). The increasing shortage of radiology professionals, exacerbated by the growing volume of imaging, is leading to critical levels of stress and burnout among staff (RCR, 2022) and causing delays and disparities in the delivery of critical care (Rimmer, 2017). All this necessitates a greater dependence on trainees and non-radiology physicians as well as teleradiology services who often lack access to the full clinical records to fill the gaps in image interpretation (Huang et al., 2023; Limb, 2022).

---

[*] Joint first authors.
[°] Joint senior authors.
Corresponding: {shruthi.bannur, kenza.bouzid, jaalvare, sthyland}@microsoft.com.

A pivotal task in radiology practice is the automated generation of a free-text report based on medical images (Liu et al., 2019). Research and development into the use of artificial intelligence (AI) to automatically generate such narrative-style radiology reports from images suggest significant potential for enhancing operational efficiency, reducing radiologist workloads, and improving the quality and standardisation of patient care (Huang et al., 2023; Liu et al., 2019; Yildirim et al., 2024; Yu et al., 2023b). AI capabilities in automatically generating draft radiology reports within seconds of image acquisition may support radiologists in better managing high case volumes, mitigating exhaustion and circumventing the constraints of human resources (Huang et al., 2023). Furthermore, rapid image interpretation by AI could reduce unnecessary variations in treatment, decrease the number of patients needing to be recalled after discharge, and support the prioritisation of urgent cases by highlighting critical conditions that require immediate clinical attention (Huang et al., 2023).

Consequently, the ability to generate free-text, narrative-style reports from radiology images has become subject to increasing research interest (Sloan et al., 2024; Zhou et al., 2024; Yang et al., 2024; Tu et al., 2024; Hyland et al., 2023; Jeong et al., 2023; Tanida et al., 2023; Chen et al., 2024; Müller et al., 2024; Wang et al., 2023; Li et al., 2023). The open-ended nature of the report generation task requires the model to describe not simply the presence or absence of findings, but more subtle aspects such as severity and extent, location, texture or other qualities, and whether the finding has progressed or resolved. Thus, report generation is a challenging task for multimodal AI models in terms of both image understanding and language generation.

For a draft radiology report to be useful, it must: (i) replicate what the radiologist would have written, without hallucinations or omissions, and (ii) be easy to verify as such. We draw on work outlining the role of reporting *context* in the generation of reports (Nguyen et al., 2023; Bannur et al., 2023) – namely that additional inputs to the AI model, such as the *Indication* and the prior study, are required to faithfully reproduce a report. We then extend the task of report generation to *additionally* require the model to *ground* each described finding in the image by generating image-level localisation annotations, such as bounding boxes. We call this task **grounded report generation**, inspired by Moor et al. (2023). The ability to ground report findings or phrases within the relevant region in medical images has been described to play a significant role: (i) in assisting image understanding and radiological diagnosis (Chen et al., 2023; Yildirim et al., 2024; Zou et al., 2024); and (ii) for verifying the correctness of AI text outputs (Bernstein et al., 2023) – a key property to support the integration of automated report drafting systems in radiology workflows.

User research with radiologists and clinicians (Yildirim et al., 2024) demonstrates that although radiologists are capable of identifying relevant findings on an image via text location description alone (e.g., left lung consolidation), this can be more difficult when findings are small or overlapping (e.g., small pneumothorax, mass behind the heart); with more complex imaging; and when assessing images outside the reporter's core area of expertise. Grounded reporting may also have utility for non-radiology clinicians, where image grounding can support comprehension and a deeper engagement with the image beyond the text report (Yildirim et al., 2024); and to improve communication with patients when reviewing image findings (Zou et al., 2024).

Grounded reporting differs from the existing task of medical phrase grounding (MPG) (Müller et al., 2024; Ichinose et al., 2023; Zou et al., 2024; Boecking et al., 2022) in that MPG aims to ground a *specified* finding or phrase, typically assumed present within the image. A grounded report is a description of *all* findings in an image with accompanying localisation. A variant of this task was explored in Tanida et al. (2023), where the model first located *anatomical* regions before generating region-level descriptions. To overcome the many-to-many challenge faced by Tanida et al. (2023), where a single sentence in a report can describe multiple findings and hence several regions, we design a dataset such that each sentence describes at most a single finding, enabling precise localisation.

Context beyond a single image plays a significant role in the contents of a radiology report, influencing both the interpretation of the image and communicative choices in the reporting itself. Hence, in this work we generate chest X-ray (CXR) reports using: the current frontal image, the current lateral image, the prior frontal image and prior report, and the *Indication*, *Technique*, and *Comparison* sections of the current study.

Selective reporting of findings is mediated by the *Indication* (Nguyen et al., 2023) for the study – a report should 'answer' any question it poses – which further provides health context on the patient (Yapp et al., 2022).

Empirically, providing the *Indication* to the model improves the quality of generated reports (Dalla Serra et al., 2022; Hyland et al., 2023) and has become more commonplace (Tu et al., 2024; Chaves et al., 2024; Yang et al., 2024). Similarly, comparison to previous imaging studies is crucial for tracking the development of disease or impact of treatment, and references to prior studies are frequent in radiology reporting (Aideyan et al., 1995; Bannur et al., 2023). Such references can be removed to reduce hallucinations when prior studies are not available (Ramesh et al., 2022; Chaves et al., 2024; Nguyen et al., 2023), or used in conjunction with prior images to enable descriptions of change (Bannur et al., 2023; Dalla Serra et al., 2023; Zhu et al., 2023).

The lateral view in a CXR study provides complementary information to frontal (AP/PA) views. It is required to identify findings like vertebral compression fractures or small pleural effusions behind the diaphragm, and can assist in the detection and differentiation of conditions such as lung nodules, masses, and certain types of pneumonia. Incorporating the lateral view has been demonstrated to improve automated report generation (Liu et al., 2024; Lee et al., 2023; Mondal et al., 2023; Yang et al., 2020; Yuan et al., 2019).

The *Technique* and *Comparison* sections provide additional context for the circumstances of the study the contents of the report: *Technique* can include indicators of patient positioning (e.g. supine, upright) and *Comparison* is informative for whether the radiologist consulted prior studies. We show these factors can make a significant impact on the accuracy of the generated report.

The new task of grounded report generation needs a novel evaluation approach. Inspired by factuality-based methods (Min et al., 2023; Schumacher et al., 2023; Xie et al., 2023), we propose an evaluation framework named RadFact. Radiology-specific metrics typically provide complementary information to *n*-gram based approaches such as BLEU or ROUGE by prioritising radiology-specific information, extracted via specialised models such as CheXbert (Smit et al., 2020; Irvin et al., 2019) or RadGraph (Jain et al., 2021; Yu et al., 2023b; Delbrouck et al., 2022). The CheXpert findings classes (Irvin et al., 2019) are biased towards intensive care settings and confound radiological findings (e.g., consolidation) with clinical diagnoses (e.g., pneumonia). Further, the use of specialised models limits their application to 'in-distribution' datasets where model behaviour is more stable (Yang et al., 2024). More recently, approaches leveraging LLMs have been proposed. Owing to their broad pre-training, LLMs are expected to generalise well to novel reporting styles. Leveraging the ReXVal dataset (Yu et al., 2023a), methods such as CheXprompt and LLM-RadJudge use LLMs to estimate the number of errors per report (Chaves et al., 2024; Wang et al., 2024).

Building on the observation that GPT-4 exhibits strong logical reasoning capabilities in radiology (Liu et al., 2023c), RadFact leverages LLMs to ascertain the factuality of *each* sentence in a generated report, given the reference ground truth. This provides sentence-level information on errors, allowing us to estimate the review and editing burden of the report in a draft-then-review setting. Further, RadFact handles evaluation of *grounded* reports by using logical evidence to match generated and ground-truth annotations.

In summary, the contributions of this work are:

1. We extend the report generation task with more useful outputs – grounded report generation – and more comprehensive inputs (lateral view, prior frontal, prior report, *Indication*, *Technique*, *Comparison* sections), and demonstrate their utility.

2. We propose a novel metric suite leveraging the logical inference capabilities of LLMs, RadFact, for report generation both with and without grounding.

3. We construct and analyse MAIRA-2, a CXR-specialised multimodal model capable of generating grounded and non-grounded reports. MAIRA-2 establishes a new state of the art on MIMIC-CXR findings generation.

2 Methods

2.1 Grounded radiology reporting

We define a grounded report as a list of sentences, (1) each associated with *zero or more spatial image annotations* and (2) describing at most a single finding from an image, as shown in Figure 1.

Spatial annotations indicate the region of the image pertaining to the described finding, and should be as specific as possible while containing the finding. Sentences describing non-findings ('No pneumothorax'), regions of normality ('Lungs are clear'), or abnormal findings without specific location ('Diffuse disease') do not require such annotations. In this work, we use bounding boxes as spatial annotations. Bounding boxes are commonly used to localise findings on CXRs (Nguyen et al., 2022; Wang et al., 2017; Boecking et al., 2022; Müller et al., 2024) and are easier to annotate than full segmentation masks.

Reports frequently include sentences with multiple radiological findings, or mentions of extrinsic information that cannot be objectively inferred by a model from the data available at reporting time (e.g., differential diagnoses, recommendations for follow-up examinations, communications with other healthcare staff). To make grounded reporting a well-defined task and enable precise association of spatial annotations with individual findings, we generate reports as lists of sentences describing individual positive or negative findings. Then, each sentence with a finding can be associated with spatial annotations.

2.2 Data

Table 1: Datasets used in the training and evaluation of MAIRA-2. For report generation tasks (findings generation and grounded reporting), a sample consists of at least one image, a findings section, and other report sections. For phrase grounding, a sample is an image with a corresponding single phrase and one or more bounding boxes. `FindGen` = findings generation, `GroundRep` = grounded reporting, `PhraseGround` = phrase grounding. 'All' means all studies with a *Findings* section. Statistics on laterals and priors are percentages of samples. Having a prior means having a prior study, including a report and a frontal image. MIMIC-CXR: Johnson et al. (2019a). MS-CXR: Boecking et al. (2022). PadChest: Bustos et al. (2020). USMix is private, with a mix of in-patient and out-patient facilities in the US. IU-Xray: Demner-Fushman et al. (2016). Datasets not used in evaluation have '–' for test set numbers. * IU-Xray has no patient information so we report study information.

Data source	Subset	Task	# Patients		# Samples		% Has Lateral		% Has Prior	
			Train	Test	Train (%)	Test	Train	Test	Train	Test
MIMIC-CXR	All	FindGen	55 218	285	158 555 (32%)	2461	60.6	45.3	64.2	88.6
	MS-CXR	PhraseGround	595	128	817 (0.2%)	176	0	0	0	0
PadChest	All	FindGen	49 029	6162	79 758 (16%)	6313	45.9	47.1	38.5	31.9
USMix	All	FindGen	118 031	–	193 652 (39%)	–	51.7	–	0	–
	GR-1	GroundRep	45 155	–	60 463 (12%)	–	48.0	–	0	–
	GR-Bench	GroundRep	8458	1199	8580 (1.7%)	1231	81.2	79.8	0	0
IU-Xray	All	FindGen	–	3198*	– –	3306	–	92.1	–	0
Total		Multi-task	222 278	–	501 825 (100%)	–	54.0	–	26.7	–

There are not yet public datasets suitable for developing a grounded reporting model[1]. For this study, we obtained a private CXR dataset (USMix) from which we derived grounded reports. We processed the narrative report text into individual sentences, as detailed in Section 2.1 and Appendix A.2, and then obtained bounding-box annotations from radiologist annotators. USMix is described in more detail below.

To enable comparison with prior work, we additionally used a set of public datasets. MIMIC-CXR (Johnson et al., 2019a) and PadChest (Bustos et al., 2020) are CXR datasets containing multiple views (frontal, lateral),

---

[1]We will release MAIRA-2 performance on a public grounded reporting benchmark dataset based on PadChest in a future version of this work.

and temporal linking enabling the use of prior study information. IU-Xray (Demner-Fushman et al., 2016) is a CXR dataset containing frontal and lateral views, without patient-level metadata, hence no prior study information. Each dataset supports possible tasks of `FindGen`: generating the *Findings* section of the report, `GroundRep`: grounded reporting as described in Section 2.1, and `PhraseGround`: generation of bounding boxes given an input phrase. The full set of datasets used in both training and evaluation is outlined in Table 1.

MIMIC-CXR (Johnson et al., 2019a) For MIMIC-CXR we extract each report's *Findings*, *Indication*, *Technique*, and *Comparison* sections following Johnson et al. (2019a). We also use the MIMIC-CXR-derived phrase grounding dataset MS-CXR (Boecking et al., 2022), which contains individual phrases from reports and associated bounding boxes for a fixed set of pathologies. We follow the official MIMIC-CXR split (Johnson et al., 2019b), with the exception of studies in MS-CXR, which are not well-distributed across the official MIMIC-CXR splits. For MS-CXR, we define a patient-level split stratified by pathology, age, and sex [2]. Studies in the MS-CXR test and validation folds are not used in training - otherwise we follow the official MIMIC-CXR split. We note that the official MIMIC-CXR test split is highly enriched for abnormal cases (Johnson et al., 2019a), hence prior studies are more common (Table 1).

PadChest (Bustos et al., 2020) The reports in the PadChest dataset are originally in abbreviated Spanish. We use the GPT-4-translated English version from the Interpret-CXR collection used in the RRG24 competition (Xu et al., 2024), which included only the *Findings* and *Impression* sections. PadChest does not have an official split, so we construct a patient-level pathology-stratified split.

USMix Our private dataset, USMix, is sourced from a set of US hospitals with a mix of in- and outpatient studies. We extract section text using GPT-4. No temporal study linkage is possible for this data source, so while we do not use prior study information, reports can contain references to prior studies. Two subsets of this dataset have been additionally annotated for grounded reporting (Section 2.1), using slightly different protocols: `GR-1` and `GR-Bench`. Protocol differences produced, for example, fewer but larger boxes per finding in `GR-1` compared to `GR-Bench`, especially for bilateral findings. We consider `GR-Bench` our benchmark and report test results on a held-out portion of it.

IU-Xray (Demner-Fushman et al., 2016) We use the entire IU-Xray dataset for external validation for the task of *Findings* generation. Reports in this dataset are stored in XML format with sections pre-extracted. The *Technique* section was taken from each image caption. We also process the dataset to use the same indicator for deidentified information as used in MIMIC-CXR ("__").

For all datasets, we drop studies missing the *Findings* section. Each frontal view in a study is treated independently. If there are multiple laterals available, we select one randomly. At training time, for MIMIC-CXR if there are multiple frontal images in the prior study, all pairings of current and prior frontal images are used as individual samples. For PadChest we select a prior frontal randomly.

We resized the original DICOM files isotropically with B-spline interpolation so that their shorter side was 518, min-max scaled intensities to [0, 255], and stored them as PNG files. At training time, we centre-crop images to 518×518 pixels before applying z-score normalisation with statistics (mean and variance) derived from MIMIC-CXR. We used SimpleITK for all image preprocessing operations (McCormick et al., 2014).

2.3 MAIRA-2

MAIRA-2 is built with a similar architecture to MAIRA-1 (Hyland et al., 2023), based on the LLaVA framework (Liu et al., 2023b;a). We use a re-trained version of RAD-DINO (Pérez-García et al., 2024) as the frozen image encoder, which is an 87M-parameter ViT-B (Dosovitskiy et al., 2020); the language model is initialised to the weights of Vicuna 7B v1.5 (Chiang et al., 2023), itself finetuned from Llama 2 (Touvron et al., 2023); and the adapter is a randomly initialised multilayer perceptron (MLP) with four layers. Since RAD-DINO processes images of size 518×518 into patches of size 14×14, each image produces a sequence of 1369 visual tokens. We do not use the ⟨CLS⟩ token. We additionally train a variant of MAIRA-2 using Vicuna 13B v1.5, which we call MAIRA-2 13B.

---

[2]We will release the stratified MS-CXR split on PhysioNet as a follow-up.

Figure 1: MAIRA-2 architecture, illustrating (a) encoding of model inputs and (b) grounded report generation. Based on the LLaVA framework (Liu et al., 2023b), the model can handle arbitrarily interleaved text and images, using a frozen vision encoder and training an adapter and an autoregressive language model (see architecture details in Section 2.3). For a given radiological study, the model can be presented with all or some of the following: the current study's frontal and lateral X-ray images; indication, technique, and comparison; prior study's image and report; along with a task-specific instruction.

Tokenisation for grounding   Inspired by Pix2Seq (Chen et al., 2022), UniTAB (Yang et al., 2022), and Kosmos-2 (Peng et al., 2023), MAIRA-2 represents a bounding box in terms of discretised coordinates representing the top-left and bottom-right corners on a uniform $N \times N$ grid (N is set to 100 in all our experiments). Kosmos-2 encodes each corner using a flat vocabulary with N^2 unique tokens for every possible grid location (e.g. "$\langle \texttt{loc1234} \rangle \langle \texttt{loc5678} \rangle$" for a box with corners $(0.12, 0.34)$ and $(0.56, 0.78)$), and UniTAB uses a shared vocabulary of N tokens for both horizontal and vertical coordinates (e.g. "$\langle \texttt{coord12} \rangle \langle \texttt{coord34} \rangle \langle \texttt{coord56} \rangle \langle \texttt{coord78} \rangle$" for the same example box). Because these encoding schemes offer no inductive bias for the model to learn true 2D representations, we instead choose to separately encode horizontal and vertical coordinates as disjoint sets of $N + N$ tokens, as e.g. "$\langle \texttt{x12} \rangle \langle \texttt{y34} \rangle \langle \texttt{x56} \rangle \langle \texttt{y78} \rangle$".

Sentences with spatial annotations are represented by a combination of text, box, and delimiter tokens. Each 4-tuple of box coordinate tokens is surrounded by $\langle \texttt{box} \rangle$ and $\langle \texttt{/box} \rangle$ tokens and concatenated. These are appended to the phrase text tokens, and the entire group is delimited by $\langle \texttt{obj} \rangle$ and $\langle \texttt{/obj} \rangle$. Examples of the full sequence structure can be seen in Fig. 1.

All non-text tokens are appended to the pretrained language model's vocabulary, with corresponding embeddings initialised to the mean embedding of the existing tokens, following LLaVA (Liu et al., 2023b).

Incorporating prior and lateral images   In typical CXR datasets, each radiological study contains at least one frontal image (posteroanterior or anteroposterior view), and optionally a lateral projection image, depending on local healthcare guidelines and purpose of the study. When it is possible to identify the most recent prior study, we additionally retrieve a prior frontal image and its corresponding report. More details on how the images were selected are provided in Section 2.2. All input images are fed through the same encoder and adapter module to obtain image tokens. These are then interleaved with the text tokens at specified placeholder locations. The full prompt structure is shown in Table 2.

Training   We train MAIRA-2 with a conventional autoregressive cross-entropy loss in a multitask setting on the dataset mix shown in Table 1. Each sample in a batch has a task and input-specific prompt as outlined in Table 2. Following Hyland et al. (2023), we do a single stage of training with a frozen image encoder, training the adapter and all the parameters of the LLM. We train for three epochs and use the final checkpoint in evaluations. We use the AdamW optimiser (Loshchilov & Hutter, 2019) with a global batch size of 128 across

Table 2: **Prompt structure**. As shown in Fig. 1, the language model receives a sequence of tokens obtained by concatenating the following messages, replacing placeholders indicated by {brackets}. Each image placeholder is replaced with 1369 image tokens encoded by RAD-DINO. Report section placeholders are replaced by the corresponding section from the sample, if available, otherwise 'N/A'. For samples missing the lateral view or prior study, we entirely remove that part of the prompt, avoiding references to nonexistent image views. We show here the instruction for GroundRep. For FindGen, the instruction is simply "provide a description of the findings in the radiology study."

Message type	Message
System	You are an expert radiology assistant tasked with interpreting a chest X-ray study.
Current frontal	Given the current frontal image {frontal_image_tokens}
Current lateral	the current lateral image {lateral_image_tokens}
Prior frontal	and the prior frontal image {prior_image_tokens}
Prior report	PRIOR_REPORT: {prior_report}
Instruction	provide a description of the findings in the radiology study. Each finding should be described as a self-contained plain-text sentence. If the finding is groundable, locate the finding in the current frontal chest X-ray image, with bounding boxes indicating all locations where it can be seen in the current frontal image. Otherwise, generate just the ungrounded finding without bounding boxes
Indication	INDICATION: {indication}
Technique	TECHNIQUE: {technique}
Comparison	COMPARISON: {comparison}

16 NVIDIA A100 GPUs, a cosine scheduler with a warm-up of 0.03, and a learning rate of 2×10^{-5}. In addition, we use a linear RoPE scaling factor of 1.5 in order to extend the context length of the LLM to handle up to 3 view images and additional inputs.

We retrained the image encoder, RAD-DINO (Pérez-García et al., 2024), for 106 000 iterations starting from the public ViT-B weights (Oquab et al., 2024), using a global batch size of 1280 across 32 A100 GPUs. The source datasets are the same as in Pérez-García et al. (2024), though we excluded from the training set all images used for evaluation in this manuscript. Table A.1 provides the number of images from each dataset.

2.4 RadFact: An evaluation suite for (grounded) reports

Grounded report evaluation requires an assessment of both textual quality and grounding correctness. A good report should be complete and concise, correctly describing the findings in the ground-truth report without extraneous detail or hallucinatory observations. The grounded regions should be specific to the finding in their associated sentence, and neither too large nor too small.[3]

To this end, we developed a framework called RadFact for the evaluation of model-generated radiology reports given a ground-truth report, which naturally enables evaluation of grounding annotations if present. RadFact provides a fine-grained *suite* of metrics, capturing aspects of precision and recall at both text-only and text-and-grounding levels.

Logical entailment   Inspired by approaches such as FActScore (Min et al., 2023), we leverage a model that can perform entailment verification (Sanyal et al., 2024) to classify whether a candidate sentence ('hypothesis') is logically true given a reference text ('premise'). A class of models suitable for entailment verification are LLMs (Liu et al., 2023c). In this work, we use Llama3-70B-Instruct[4] (AI@Meta, 2024) for entailment verification with ten in-context examples – we refer to this version as RadFact-Llama3 in tables, noting that different backend LLMs can produce different behaviour.

---

[3]We define this evaluation generically in terms of "regions" to encompass diverse forms of grounding, including single or multiple bounding boxes, polygons, segmentation masks, etc.

[4]From: https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

Figure 2: Illustration of RadFact. Zoomed panel (bottom) shows a single direction of evaluation, taking the model generations as logical hypotheses and the original report as premises. Here, logical precision measures the fraction of generated sentences which are entailed according to the original report. We use a LLM with task-specific prompting to classify hypotheses as entailed or not, given premises. Grounding precision is the fraction of *logically entailed*, grounded sentences whose spatial annotations are also entailed. Spatial precision is the fraction of *all* grounded sentences whose spatial annotations are also entailed, hence it is upper-bounded by grounding precision. Analogously, in the opposite direction recall can be computed by taking the original sentences as hypotheses and using the model generations as premises (top right). Here, spatial annotations are bounding boxes. A sentence can have multiple boxes (see sentence B). Spatial entailment requires a pixel precision above 0.5, e.g. at least 50% of the pixels associated with the sentence fall into a matched evidence box. In the above, sentence B's evidence comes from premises 4 and 5, hence its boxes are compared with the boxes from 4 and 5.

The task is illustrated in Fig. 2. The generated and ground-truth reports are assumed to consist of lists of sentences, each describing a single finding. In a conventional findings-generation scenario, free-text reports can first be converted into this format as described in Section 2.1 and Appendix A.2.

RadFact computes entailment in both directions, defining the following text-level metrics:

1. RadFact logical precision: the fraction of generated sentences that are entailed by the ground-truth report. This measures how truthful the model generations are, as it penalises hallucinations.

2. RadFact logical recall: the fraction of ground-truth sentences that are entailed by the generated report. This measures how complete the generated report is, as it penalises omissions.

This bidirectional approach differs from traditional factual verification approaches such as FActScore that assume a 'single' source of truth (e.g., Wikipedia), but has precedents in medical summarisation where both completeness and conciseness are important (Xie et al., 2023).

We further require the entailment verification model to provide *evidence* for its classification: this is the set of premise sentences from the reference report that support the determination of entailment (or not) for each hypothesis[5]. Evidence may be empty for logically neutral statements, which are considered not-entailed by definition. Evidence enables us to match the grounding regions from generated sentences with their (supposed) ground-truth regions.

Spatial and grounding entailment We can then define a notion of *spatial entailment* based on pixel overlap: a region is spatially entailed by its evidence region(s) if at least a given fraction of its pixel mask is contained in the evidence pixel mask.[6] This definition interprets a larger region as *more specific* than a smaller region contained within it, as the former makes stronger claims about where a finding is located. This provides for metrics on the text-and-grounding quality, analogously defining precision based on sentences from the generated report, and recall based on sentences from the ground-truth report:

1. RadFact grounding {precision, recall}: the fraction of *logically entailed* grounded sentences that are *also* spatially entailed. This tells us: which of the correctly *described* findings were also *correctly grounded*?

2. RadFact spatial {precision, recall}: the fraction of *all* grounded sentences that are *logically and spatially* entailed. This metric additionally penalises grounding incorrect sentences.

Figure 2 further illustrates how these metrics are defined. The fractions are calculated once in each direction: 'precision' scores describing the correctness of generated findings with respect to the ground-truth report, and conversely 'recall' scores indicating their completeness.

Appendix A.3 provides more detail about RadFact-Llama3, such as the system prompt and few-shot examples.

2.5 Evaluation and metrics

We supplement RadFact, and enable comparison with prior work by computing a set of commonly-reported metrics. Since MAIRA-2 generates both text and boxes, we distinguish here between evaluation of text outputs, and evaluation of boxes (grounding). To quantify variance in the model's test set performance, we perform bootstrapping and report median and 95% confidence intervals over 500 replicates.

Text-only evaluation. We employ a combination of traditional natural language generation (NLG) ('lexical') metrics and radiology-specific ('clinical') metrics. For lexical metrics, we use ROUGE-L (Lin, 2004), BLEU-{1,4} (Papineni et al., 2002), and METEOR (Banerjee & Lavie, 2005). For clinical metrics, we use RadGraph-F1 (Jain et al., 2021), RG_{ER} (Delbrouck et al., 2022)[7], RadCliQ version 0 (Yu et al., 2022), and CheXbert vector similarity (Smit et al., 2020; Yu et al., 2023b)[8], as well as macro- and micro-averaged F1 scores for CheXpert classes (Irvin et al., 2019) based on the CheXbert classifier (Smit et al., 2020). We further report CheXprompt scores, which uses GPT-4 to estimate the number of errors in a generated report. Following Chaves et al. (2024) we report the mean errors per report, as well as the percentage of error-free reports, distinguishing between any errors, and significant errors.

Grounding-only evaluation. To evaluate bounding boxes independently of text generation, we employ a box-completion approach similar to Peng et al. (2023). The model is conditioned on the prompt and the grounded report up to and including the target phrase and the first ⟨box⟩ token, and is allowed to generate boxes until a closing ⟨/obj⟩ token is produced. We do this for every grounded phrase over all reports in the dataset, then compute spatial overlap metrics between the pixel masks of the completed boxes and of the

---

[5]RadFact does not require a one-to-one mapping between generated and reference sentences, and there can be several pieces of evidence to support a logical inference. For example, the sentence 'bilateral pleural effusions' implies both 'left pleural effusion' and 'right pleural effusion' simultaneously, hence it can be used as evidence for either. Conversely, *both* 'left plerual effusion' and 'right pleural effusion' are required to support the conclusion of 'bilateral pleural effusions'.

[6]This pixel-precision threshold is set to 0.5 in our implementation with multiple boxes as the form of grounding, but could be adjusted, e.g., for finer-grained segmentation masks.

[7]RG_{ER} is implemented as F1RadGraph with reward=partial by https://pypi.org/project/radgraph/.

[8]For RadGraph-F1, RadCliQ and CheXbert vector similarity, we use https://github.com/rajpurkarlab/CXR-Report-Metric.

Table 3: **Grounded reporting performance on the test fold of `GR-Bench`.** We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Metric	MAIRA-2 7B		MAIRA-2 13B	
Lexical: ROUGE-L	58.2 [56.7, 59.8]		59.4 [57.8, 61.0]	
RadFact:	Precision	Recall	Precision	Recall
Logical	73.5 [72.2, 74.9]	72.4 [71.0, 73.8]	74.8 [73.5, 76.3]	73.3 [72.0, 74.7]
Spatial	32.1 [29.4, 34.5]	33.7 [31.2, 36.2]	35.0 [32.4, 37.8]	36.9 [34.5, 39.5]
Grounding	68.2 [64.7, 71.7]	92.2 [89.8, 94.4]	68.8 [65.3, 71.9]	91.1 [89.1, 93.1]
Clinical:				
RadGraph-F_1	54.2 [52.5, 55.9]		55.9 [54.1, 57.6]	
RG_{ER}	56.9 [55.3, 58.5]		58.4 [56.7, 60.1]	
RadCliQ (↓)	1.63 [1.55, 1.70]		1.56 [1.49, 1.64]	
CheXbert Macro F_1-14	40.9 [35.9, 47.1]		45.9 [40.1, 52.4]	
CheXbert Micro F_1-14	60.2 [57.5, 62.5]		61.4 [59.0, 63.8]	
Phrase grounding:	Precision	Recall	Precision	Recall
Box-completion	68.4 [67.2, 69.7]	84.6 [83.7, 85.5]	70.2 [68.8, 71.7]	86.2 [85.4, 87.1]

respective ground-truth boxes. Note that `RadFact` quantifies grounding on the sentence level in a binary fashion, whereas this complementary pixel-level evaluation measures the quality of the boxes in isolation.

3 Results

3.1 MAIRA-2 can generate grounded reports

To the best of our knowledge, MAIRA-2 is the first model that both generates full report sections and grounds each detected finding in the image, and thus serves as a baseline for future work on this task. We report the test performance of MAIRA-2 for grounded report generation on the `GR-Bench` dataset in Table 3. `RadFact` logical scores are consistently above 70%, indicating a low rate of both omissions and hallucinations. Conditional on the model first generating a correct sentence (`RadFact` grounding precision and box-completion precision), 68%-70% of such sentences are correctly grounded. The drop in `RadFact` spatial however demonstrates that the model generates boxes that are either 1) incorrect, or 2) associated with incorrect sentences. The high `RadFact` grounding *recall* indicates that the model is reliably generating boxes which contain ground truth boxes. Scaling to 13B provides modest improvements in text quality (via `RadFact` logical scores, and other clinical metrics) and a more significant improvement on localisation-based metrics.

To better understand and demonstrate the behaviour of MAIRA-2 in grounded reporting, we selected examples for qualitative review with a radiologist, shown in Figure 3 and Appendix C.

Phrase grounding on MS-CXR Because there are no previously published results for grounded reporting, MAIRA-2 was also evaluated on the related task of phrase grounding, for which public baselines exist. Phrase grounding here means generating a set of bounding boxes given an image and an input phrase, such as 'left retrocardiac opacity'. We compare against MedRPG (Chen et al., 2023), ChEX (Müller et al., 2024), and TransVG (Deng et al., 2021). Compared to MAIRA-2, these baselines directly regress bounding box coordinates using MLP heads. MedRPG additionally employs a combination of contrastive and attention losses to better align image- and text-features. Similarly, the phrase grounding in ChEX benefits from the synergies of multitask training, combining report generation and localisation tasks.

Table 4 presents the mean intersection over union (mIoU) of pixel masks from generated vs ground-truth boxes on our test split of the MS-CXR dataset (Boecking et al., 2022). Note that Chen et al. (2023) and Müller et al. (2024) used different custom splits of MS-CXR. To enable fair comparison, we therefore report comparative results on the intersections of our test set with their respective test subsets. The 95% confidence intervals for

Figure 3: **A manually-selected qualitative example of MAIRA-2 output on GR-Bench**. This 3-part figure shows MAIRA-2 model inputs (top); the MAIRA-2 phrase outputs vis-à-vis the reference text (middle); and grounding boxes for the MAIRA-2 phrases on the current frontal image alongside their alignment with reference boxes (bottom). In this example, all generated MAIRA-2 phrases were evidenced by the reference text (RadFact logical precision: 1.0). In a radiologist review, we find two missed findings: "There is a calcified granuloma in the right lower lung." and "A slight compression fracture of t1 is noted.", which can only be seen on the lateral view. RadFact further counts finding 11 as missed, bringing logical recall to 0.75. Reviewing the reference findings, radiologists pointed out that the compression fracture is on L1 vertebra in the image, suggesting a potential typo in the reference text. Concerns were also raised that small fracture cases may not always be reported and could be missed in training data. Although the compression fracture was not detected, MAIRA-2 correctly outputs the "degenerative changes of the spine" that are always better seen on the lateral view. For image grounding, no boxes were generated for missed findings 2 and 7. While finding 11 ("There is calcification within the knob") was also not logically entailed according to RadFact, the model did correctly generate a separate box around the aortic knob when grounding finding H ("There is uncoiling of the aorta with calcification").

11

Table 4: **Phrase grounding performance (mIoU) on MS-CXR**. MedRPG (Chen et al., 2023) reports performance on 20% of the single-box cases from MS-CXR (approx. 178 phrases, 162 images), whereas ChEX (Müller et al., 2024) included only samples in the official MIMIC-CXR validation and test splits (196 phrases, 169 images). Because the final MAIRA-2 model was trained with a part of MS-CXR, we report results on the intersections of our new held-out test split (176 phrases, 155 images) and each of the splits from MedRPG (138 phrases, 124 images) and ChEX (30 samples, 24 images), respectively. Results for TransVG (Deng et al., 2021) are quoted here from the comparisons originally reported for MedRPG and ChEX.

Model	Single-box only	In MIMIC-CXR val./test	Test split
	$(n \approx 178)$	$(n = 196)$	–
MedRPG	59.37	–	–
ChEX	–	46.51 [44.68, 50.36]	–
TransVG	58.91	53.51 [50.51, 56.51]	–
	$(n = 138)$	$(n = 30)$	$(n = 176)$
MAIRA-2	57.78 [54.00, 61.40]	57.56 [49.44, 65.17]	54.70 [51.35, 58.21]
MAIRA-2 13B	61.88 [58.46, 65.32]	60.88 [53.41, 67.71]	59.29 [56.34, 62.31]

ChEX and TransVG were approximated assuming a normal distribution based on the bootstrapped standard deviation reported by Müller et al. (2024).

On the phrase grounding task, MAIRA-2 achieves competitive performance against baselines developed specifically for phrase grounding (MedRPG and TransVG) and appears to strongly outperform the multi-task ChEX model. Moreover, the larger MAIRA-2 13B version further improves grounding results.

3.2 MAIRA-2 is state-of-the-art on findings generation

MAIRA-2 is designed and trained to handle both grounded or non-grounded report generation. Table 5 shows its performance on the MIMIC-CXR test set using both RadFact and a range of commonly reported metrics outlined in Section 2.5. We compare to the closest prior state of the art, restricted to models evaluated for *Findings* generation, namely Med-PaLM M (Tu et al., 2024), LLaVA-Rad (Chaves et al., 2024), MedVersa (Zhou et al., 2024), and MAIRA-1 (Hyland et al., 2023). Since many of these models are not publicly available, we present their evaluation results as originally reported, noting that the test sets are slightly different. For MAIRA-1, we obtained the model generations on the MIMIC-CXR test set in order to run RadFact and CheXprompt.

Table 5 shows that MAIRA-2 outperforms or matches all prior approaches across all metrics, with MAIRA-2 13B providing further improvements. The impact on lexical metrics is most significant, where MAIRA-2 improves on existing scores by 14% to 27%. On existing clinical metrics, significant improvement is observed on the RadGraph-derived RadGraph-F_1 and RG_{ER}, on the CheXbert vector score, and on CheXbert 14-class F_1, using both micro and macro averaging. For RadCliQ, MAIRA-2 and MedVersa have overlapping confidence intervals, but a significant improvement is seen with MAIRA-2 13B. In the following sections, we explore the features of MAIRA-2 which result in these improvements.

With RadFact, we see again an improvement from MAIRA-1 to MAIRA-2 and its 13B variant, in agreement with other metrics. What RadFact additionally reveals is that in *absolute* terms, models continue to make errors, with only 55.6% of sentences generated by MAIRA-2 13B confirmed (as per the reference report) to be true. We show qualitative examples of MAIRA-2 generations on MIMIC-CXR in Appendix C.3.

Although there is no prior work demonstrating findings generation performance on PadChest in English, in Table 6 we show the performance on MAIRA-2 to enable future comparison. External performance on IU-Xray, shown in Appendix B.1 and Table B.1, demonstrates that MAIRA-2 can generalise to novel reporting scenarios, achieving RadFact logical precision and recall of 71% and 68% respectively.

Table 5: **Findings generation performance on the official MIMIC-CXR test split**. [†] means numbers were taken from prior work, except for `RadFact` and CheXprompt for MAIRA-1(Hyland et al., 2023). Med-PaLM M: Tu et al. (2024). LLaVA-Rad: Chaves et al. (2024). MedVersa: Zhou et al. (2024). We report median and 95% confidence intervals based on 500 bootstrap samples. **Bold** indicates best performance for that metric, or overlapping CIs with best. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Metric	MAIRA-1	Med-PaLM M[†]	LLaVA-Rad[†]	MedVersa[†]	MAIRA-2	MAIRA-2 13B
Lexical:						
ROUGE-L	28.9 [28.4, 29.4]	27.29	30.6	–	**38.4** [37.8, 39.1]	**39.1** [38.5, 39.7]
BLEU-1	39.2 [38.7, 39.8]	32.41	38.1	–	46.5 [45.8, 47.2]	**47.9** [47.2, 48.7]
BLEU-4	14.2 [13.7, 14.7]	11.31	15.4	17.8 [17.2, 18.4]	23.4 [22.9, 24.0]	**24.3** [23.7, 24.9]
METEOR	33.3 [32.8, 33.8]	–	–	–	41.9 [41.3, 42.6]	**43.0** [42.4, 43.6]
RadFact:						
Logical precision	48.3 [47.3, 49.4]	–	–	–	52.5 [51.6, 53.5]	**55.6** [54.6, 56.7]
Logical recall	47.2 [46.3, 48.2]	–	–	–	48.6 [47.7, 49.6]	**51.5** [50.6, 52.5]
Clinical:						
RadGraph-F1	24.3 [23.7, 24.8]	26.71	29.4	28.0 [27.3, 28.7]	34.6 [33.9, 35.4]	**35.9** [35.6, 36.6]
RG$_{ER}$	29.6 [29.0, 30.2]	–	–	–	39.7 [38.9, 40.4]	**40.9** [40.3, 41.6]
RadCliQ (↓)	3.10 [3.07, 3.14]	–	–	2.71 [2.66, 2.75]	2.64 [2.61, 2.68]	**2.59** [2.56, 2.63]
CheXbert vector	44.0 [43.1, 44.9]	–	–	46.4 [45.5, 47.4]	50.6 [49.7, 51.5]	**51.3** [51.0, 52.1]
CheXprompt						
Mean significant errors (↓)	2.41 [2.35, 2.46]	–	2.25	–	2.22 [2.16, 2.27]	**2.07** [2.02, 2.12]
Mean errors (↓)	2.49 [2.44, 2.54]	–	2.95	–	2.31 [2.26, 2.36]	**2.16** [2.12, 2.22]
% Significant error free	4.65 [3.88, 5.55]	–	6.79	–	**7.23** [6.14, 8.33]	**8.70** [7.72, 9.87]
% Error free	3.13 [2.43, 3.86]	–	2.58	–	**5.00** [4.16, 5.87]	**5.99** [5.08, 7.01]
CheXpert F1, uncertain as negative:						
Macro-F1-14	38.6 [37.1, 40.1]	39.83	39.5	–	**42.7** [40.9, 44.4]	**43.9** [42.2, 45.7]
Micro-F1-14	55.7 [54.7, 56.8]	53.56	57.3	–	**58.5** [57.3, 59.6]	**59.0** [57.8, 60.2]
Macro-F1-5	47.7 [45.6, 49.5]	**51.60**	47.7	–	51.5 [49.3, 53.5]	51.7 [49.5, 53.8]
Micro-F1-5	56.0 [54.5, 57.5]	**57.88**	57.4	–	58.9 [57.4, 60.5]	59.1 [57.5, 60.7]

Table 6: **Findings generation performance on PadChest test split.** We use a version of the dataset which has been translated to English, and defined our own test split. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Model	ROUGE-L	BLEU-4	CheXbert Macro F_1-14	RadCliQ (↓)	RadFact Logical Precision	Recall
MAIRA-2	28.2 [27.5, 29.0]	7.7 [7.3, 8.1]	34.8 [32.5, 37.1]	2.76 [2.73, 2.78]	57.2 [56.1, 58.4]	46.3 [45.3, 47.4]
MAIRA-2 13B	29.4 [28.6, 30.1]	9.3 [8.9, 9.8]	35.9 [32.5, 37.1]	2.72 [2.69, 2.74]	57.5 [56.4, 58.6]	48.9 [47.8, 49.8]

3.3 Synergy between findings generation and grounded reporting training

MAIRA-2 is a multitask model optimised for both `FindGen` and `GroundRep` tasks. Since `GroundRep` is based on `FindGen`, we expect positive transfer between these tasks. Here we compare the performance of MAIRA-2 (7B) to models trained *only* on the task of interest, dropping either `FindGen` and evaluating on `GroundRep` (Figure 4 and Table B.2), or dropping `GroundRep`[9] and evaluating on `FindGen` (Figure 5 and Table B.3).

Figure 4: Impact of dropping the `FindGen` task (MIMIC-CXR, PadChest, USMix) on `GR-Bench` grounded reporting. First row: impact on the text-only metrics; second row: impact on grounding metrics. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. Tabular representation of this plot is available in Table B.2.

Figure 5: Impact of dropping the grounding task (`GR-1`, `GR-Bench`, and MS-CXR) on MIMIC-CXR *Findings* generation test set. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. Tabular representation of this plot is available in Table B.3.

Figure 4 shows the impact of omitting `FindGen` task from MAIRA-2 training in terms of text (top row) and box (bottom row) metrics. We find that dropping `FindGen` task results in a significant drop in all text metrics, suggesting a positive transfer between `FindGen` and `GroundRep` on the quality and clinical factuality of the generated grounded report phrases. In particular, we notice a very large decrease (-52.07%) in Macro F_1-14 when dropping `FindGen`, indicating that MAIRA-2 grounded reports identify the presence or absence of the 14 CheXpert findings more accurately when the model is trained jointly on `FindGen`. Additionally, we see a substantial decrease in `RadFact` logical precision (-6.25%) and recall (-10.36%), indicating that the model trained without `FindGen` is generating more hallucinations and omissions. This may also explain the *increase* in `RadFact` *grounding* precision (+8.94%) when we drop `FindGen`– the model is generating fewer logically entailed sentences, but those which it generates are grounded correctly more often.

While training on `FindGen` seems to improve `GroundRep` performance, we do not observe the reverse: training with `GroundRep` does not appear to benefit the `FindGen` task. Figure 5 indicates limited impact with most metrics showing overlapping confidence intervals.

3.4 Additional inputs reduce hallucinations and increase clinical accuracy for report generation

In this section, we study the impact of additional inputs used by MAIRA-2, namely the lateral view, the prior study (prior frontal and prior report), the *Technique* section, and the *Comparison* section. We do not explore dropping the *Indication* section here as its importance is already well-established (Nguyen et al., 2023; Hyland et al., 2023). We categorise these inputs along two dimensions: (i) inputs that are related to the

---

[9]We also drop `PhraseGround` when we drop `GroundRep`, to remove all grounding information during training.

temporal nature of reporting, namely the prior image and report (collectively referred to as the prior study) and the *Comparison* section; and (ii) inputs relating to multiple view types collected in a single imaging study, namely the lateral image and *Technique* section.

We perform two types of ablations on the MAIRA-2 model (7B): training ablations, wherein a model is trained and evaluated without a subset of inputs, referred to as 'Train:No <view> No <section>'; and inference ablations, dropping those inputs at inference time only 'Infer:No <view> No <section>'. For inference-time ablations, when dropping any subset of views from a sample, its prompt is constructed as if the input view was missing, and the content of dropped sections is replaced by the string 'N/A'. Further ablation experiments are discussed in Appendix B.3.

These analyses are performed on the MIMIC-CXR findings generation task, as this is a public benchmark containing linkable prior images and reports, laterals, and all the relevant report sections.

Figure 6: Impact of dropping the prior study and comparison at training and inference times 'Train:' and during inference only 'Infer:' on MIMIC `FindGen` for the 88.6% test subset that have a *Prior* ($n = 2181$). *%Comparison mentions* is estimated using Llama3-70B. The dashed line indicates the frequency of comparison mentions (91.84%) in the ground-truth reports in the same data subset, for reference. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. Tabular representation of this plot is available in Table B.4.

Inputs containing temporal information Figure 6 shows training- and inference-time ablations dropping both the prior study and the comparison section, on the subset of the MIMIC-CXR test set with prior images (n=2,181). As an additional metric, we use `Llama3-70B-Instruct` to determine whether a given report mentions temporal comparisons (see details in Appendix B.3), referred to as *%Comparison mentions*. In the absence of a prior study, *%Comparison mentions* should be close to zero. Dropping the prior study and comparison information during training 'Train:No *Prior* No *Comp*' produces a significant drop across all metrics compared to MAIRA-2 baseline, with ROUGE-L dropping by 11.7%, Macro F_1-14 by 10.1%, `RadFact` logical precision by 3.8%, `RadFact` logical recall by 8.0% and RadCliQ getting worse by 9.5% . The effect is larger when this information is dropped only at inference time 'Infer:No *Prior* No *Comp*', indicating that MAIRA-2 is effectively learning to use these inputs. Specifically, dropping the prior study and comparison section at inference time causes ROUGE-L to drop by 28.9%, Macro F_1-14 by 18.1%, `RadFact` logical precision by 13.5%, `RadFact` logical recall by 16.7% and RadCliQ to increase by 20.4%. The model trained without the prior study or comparison section hallucinates mentions of comparisons approximately 75% of the time, close to the background rate in this dataset (dashed line). Having trained with the prior study and comparison section significantly reduces such hallucinations, dropping by 49%. This further indicates the utility of training on more inputs to prevent model hallucination, since both the comparison section and the prior study contribute to the model's ability to infer whether comparison text should be generated. Additional piece-wise ablations in Appendix B.3 show that dropping the prior study has a larger effect on clinical metrics such as Macro F_1-14 and dropping the comparison has a large effect on lexical metrics as well as on clinical metrics.

15

Figure 7: Impact of dropping the lateral view and the technique section at training and inference times 'Train:' and during inference only 'Infer:' on MIMIC *Findings* generation for the 30.6% test subset that have a *Lateral* view (n=1,116). The dashed line indicates the frequency of lateral mentions (35.57%) in the ground-truth reports in the same data subset, for reference. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. Tabular representation of this plot is available in Table B.5.

Inputs related to multi-view studies Figure 7 shows the result of training- and inference-time ablations when dropping both the lateral view and the technique section, on the subset of MIMIC-CXR test studies that have a lateral view (n=1116, 30.6%). Analogously to our study on temporal information, we quantify *mentions* of the lateral view in the *Findings* section, to measure whether the model is sensitive to the presence of the lateral view[10]. *%Lateral mentions* is estimated using regular expressions (cf. Listing 6 in Appendix) to identify any explicit (e.g., AP and *lateral* views of the chest) or implicit (e.g., Chest *two* views) mentions of the lateral view. Among the 1,116 multi-view studies, 35.57% reports have at least one lateral mention in the ground truth. We analyse the impact of the lateral and the technique section separately in Appendix B.3.

When we drop the lateral and the technique section during training 'Train:No *Lat* No *Tech*', we only notice significant changes in RadCliQ (+6.4%) and ROUGE-L (-8.91%) suggesting that dropping the technique and the lateral view at training time mostly hurts lexical performance in this subset since `RadFact` and Macro F_1-14 are unchanged. This ablated model generates hallucinatory lateral mentions 36.1% of the time, while having no access to the lateral view or *Technique* section – approximately equal to the rate of lateral mentions in this subset (dashed line). On the other hand, MAIRA-2 generates dramatically fewer lateral hallucinations when we drop the lateral and the technique section at inference-time 'Infer:No *Lat* No *Tech*' with only 5.1% lateral references compared to 39.6% when the lateral and the technique are provided. Additionally, we notice a drop in lexical metrics where ROUGE-L falls down to 36.8 (-22%). The clinical metrics also decline; Macro F_1-14 is lower by 9.27%, RadCliQ increases by 4.15%, and `RadFact` phrase precision and recall decrease by 4.1% and 10.2%, respectively. This indicates that MAIRA-2 relies on the lateral view to identify the presence or absence of certain pathologies in multi-view studies. In particular, pleural effusion[11] F_1 score drops from 71.4 [66.6, 75.0] to 64.7 [59.9, 69.5]. Hence, we can conclude that using the lateral view, along with the technique section, reduces "lateral hallucinations" and improves clinical accuracy. These additional inputs prevent the model from relying on superficial cues in the training data. Instead, it enables MAIRA-2 to learn the underlying semantics of multi-view studies where the frontal and the lateral are complementary to make more comprehensive diagnoses.

---

[10]Dropping the lateral view alone is not sufficient to evaluate how well MAIRA-2 leverages laterals since it can use the *Technique* section to infer the presence of a lateral view.

[11]Lateral views in chest X-rays are particularly helpful to identify small pleural effusions behind the diaphragm, among other pathologies.

4   Conclusion

Grounded radiology report generation is a novel task that requires a model to generate image-level localisations for each finding that can be localised within the image. This enables novel uses of automatically generated reports, such as potentially more rapid verification of generated findings and use by non-radiologist clinicians, or even patients. In this work we have focused on the technical aspects of this new task to demonstrate its feasibility, leading to the development of `RadFact` and construction of MAIRA-2.

MAIRA-2 is a large multimodal model making use of the radiology-specialised RAD-DINO image encoder and the open Vicuna-1.5 large language model, in either 7B or 13B sizes. MAIRA-2 improves significantly upon the state of the art in findings generation on MIMIC-CXR owing to its more comprehensive set of inputs. Tailored to the CXR setting, MAIRA-2 leverages the current frontal and lateral views, the prior study (frontal image and full report), the *Indication* for the current study, as well as the *Technique* and *Comparison* sections. In an array of ablations, we have demonstrated the roles of these additional inputs in reducing hallucinations and improving clinical accuracy.

Our proposed evaluation framework, `RadFact`, allows for a more nuanced view of automated reporting. `RadFact` targets the core objective of evaluation in report generation: to pinpoint the errors made by the model. Using the generalisation capabilities and reasoning faculties of LLMs, `RadFact` does not rely on a fixed set of finding categories or a model which is specialised to a certain reporting style, instead operating via more flexible logical inference. Further, `RadFact` provides for sentence-level granularity on model errors, and naturally supports both grounded and non-grounded reporting. We share code for `RadFact` at https://github.com/microsoft/RadFact.

One limitation of this work is the absence of prior studies in our grounded reporting dataset, preventing an understanding of the importance of prior study information on grounding specifically. Our ablations also indicate that the model may not be using additional imaging information to the fullest extent, instead exploiting shortcuts available in the report sections used as inputs. Other methods to incorporate additional imaging information may prove superior to our token concatenation approach. Finally, we acknowledge that `RadFact` does not distinguish between the *nature* of errors beyond factuality, relying on strict logical entailment. This means some errors may be more or less clinically significant, and 'partial errors' are penalised (for example, correctly describing the presence of a pneumothorax, but not that it has improved). By open-sourcing `RadFact`, we support further improvements to enable better evaluation standards on the task of radiology report generation.

Overall we have demonstrated that grounded radiology reporting is possible with MAIRA-2. Although performance in automated report generation continues to improve – and we establish a new state-of-the-art on MIMIC-CXR with this work – metrics to date, including `RadFact`, indicate a gap between model performance and that which will be required to realise such systems in practice. The addition of grounding is a step towards real clinical impact in automated radiology report generation.

Acknowledgements

References

Uwa O Aideyan, Kevin Berbaum, and Wilbur L Smith. Influence of prior radiologic information on the interpretation of radiographic examinations. *Academic Radiology*, 2(3):205–208, 1995.

AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, June 2005. URL https://aclanthology.org/W05-0909.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Michael H Bernstein, Michael K Atalay, Elizabeth H Dibble, Aaron WP Maxwell, Adib R Karam, Saurabh Agarwal, Robert C Ward, Terrance T Healey, and Grayson L Baird. Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology*, 33(11):8263–8269, 2023.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez Valle, Hoifung Poon, and Ozan Oktay. MS-CXR: Making the most of text semantics to improve biomedical vision-language processing (version 0.1), 2022. URL https://physionet.org/content/ms-cxr/0.1/.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=e42KbIw6Wb.

Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, and Huazhu Fu. Medical phrase grounding with region-phrase context contrastive alignment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, LNCS, pp. 371–381, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43990-2_35.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 615–624, 2022.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. Controllable chest x-ray report generation from longitudinal representations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360. ACL, December 2022. doi: 10.18653/v1/2022.findings-emnlp.319.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1769–1779, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Deng_TransVG_End-to-End_Visual_Grounding_With_Transformers_ICCV_2021_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

Chanel Fischetti, Param Bhatter, Emily Frisch, Amreet Sidhu, Mohammad Helmy, Matt Lungren, and Erik Duhaime. The evolving importance of artificial intelligence and radiology in medical trainee education. *Academic Radiology*, 29:S70–S75, 2022.

Jonathan Huang, Luke Neill, Matthew Wittbrodt, David Melnick, Matthew Klug, Michael Thompson, John Bailitz, Timothy Loftus, Sanjeev Malik, Amit Phull, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA network open*, 6(10):e2336100–e2336100, 2023.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. Visual grounding of whole radiology reports for 3D CT images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224 of *LNCS*, pp. 611–621, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43904-9_59.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)*, volume 33, pp. 590–597. AAAI Press, July 2019. doi: 10.1609/aaai.v33i01.3301590.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong N. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/c8ffe9a587b126f152ed3d89a146b445-Abstract-round1.html.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest X-ray report generation. In *Medical Imaging with Deep Learning (MIDL 2023)*, 2023. URL https://openreview.net/forum?id=aZ0OuYMSMMZ.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Roger G. Mark, and Steven Horng. MIMIC-CXR database (version 2.0.0). PhysioNet, 2019a.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

S. Kalidindi and S. Gandhi. Workforce crisis in radiology in the uk and the strategies to deal with it: Is artificial intelligence the saviour? *Cureus*, 15(8):e43866, Aug 2023. doi: 10.7759/cureus.43866.

Hyungyung Lee, Da Young Lee, Wonjae Kim, Jin-Hwa Kim, Tackeun Kim, Jihang Kim, Leonard Sunwoo, and Edward Choi. Unixgen: A unified vision-language model for multi-view chest x-ray generation and report generation. *arXiv preprint arXiv:2302.12172*, 2023.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3334–3343, 2023.

Matthew Limb. Shortages of radiology and oncology staff putting cancer patients at risk, college warns. *BMJ*, 377, 2022. doi: 10.1136/bmj.o1430. URL https://www.bmj.com/content/377/bmj.o1430.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, July 2004. URL https://aclanthology.org/W04-1013.

Aohan Liu, Yuchen Guo, Jun-hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Medical Imaging*, 2024. doi: 10.1109/TMI.2024.3372638.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pp. 249–269. PMLR, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, December 2023b. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Tajdin Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya V. Nori, Matthew P. Lungren, Ozan Oktay, and Javier Alvarez-Valle. Exploring the boundaries of GPT-4 in radiology. *arXiv preprint arXiv:2310.14573*, 2023c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Matthew McCormick, Xiaoxiao Liu, Luis Ibanez, Julien Jomier, and Charles Marion. ITK: enabling reproducible research and open science. *Frontiers in Neuroinformatics*, 8, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00013. URL https://www.frontiersin.org/articles/10.3389/fninf.2014.00013.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. ACL. doi: 10.18653/v1/2023.emnlp-main.741.

Chayan Mondal, Duc-Son Pham, Tele Tan, Tom Gedeon, and Ashu Gupta. Transformers are all you need to generate automatic report from chest X-ray images. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 387–394. IEEE, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616 (7956):259–265, Apr 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4. URL https://doi.org/10.1038/s41586-023-05881-4.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. ChEX: Interactive localization and region description in chest X-rays. *arXiv preprint arXiv:2404.15770*, 2024.

Philip Müller, Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Weakly supervised object detection in chest X-rays with differentiable ROI proposal networks and soft ROI pooling. *arXiv preprint arXiv:2402.11985*, 2024.

Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. Pragmatic radiology report generation. In *Machine Learning for Health (ML4H)*, pp. 385–402. PMLR, 2023.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, July 2002. doi: 10.3115/1073083.1073135.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=lLmqxkfSIw.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. RAD-DINO: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pp. 456–473. PMLR, 2022.

RCR. Clinical radiology workforce census 2022. *The Royal College of Radiologists*, 2022. URL https://www.rcr.ac.uk/media/qs0jnfmv/rcr-census_clinical-radiology-workforce-census_2022.pdf.

Eduardo P Reis, Joselisa PQ de Paiva, Maria CB da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. BRAX, Brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022.

Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. Are machines better at complex reasoning? Unveiling human-machine inference gaps in entailment verification. *arXiv preprint arXiv:2402.03686*, 2024.

Elliot Schumacher, Daniel Rosenthal, Varun Nair, Luladay Price, Geoffrey Tso, and Anitha Kannan. Extrinsically-focused evaluation of omissions in medical summarization. *arXiv preprint arXiv:2311.08303*, 2023.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *arXiv preprint arXiv:2405.10842*, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519. ACL, November 2020. doi: 10.18653/v1/2020.emnlp-main.117.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tanida_Interactive_and_Explainable_Region-Guided_Radiology_Report_Generation_CVPR_2023_paper.html.

Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Karan Singhal, Shekoofeh Azizi, Tao Tu, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Zahra Ahmed, Sara Mahdavi, Yossi Matias, Joelle Barral, Ali Eslami, Danielle Belgrave, et al. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. *arXiv preprint arXiv:2311.18260*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, February 2024. doi: 10.1056/AIoa2300138.

UK HSA. Medical imaging: What you need to know, Sep 2022. URL https://www.gov.uk/government/publications/medical-imaging-what-you-need-to-know/medical-imaging-what-you-need-to-know--2.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11558–11567, 2023.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. LLM-RadJudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv:2404.00998*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yiqing Xie, Sheng Zhang, Hao Cheng, Zelalem Gero, Cliff Wong, Tristan Naumann, and Hoifung Poon. DocLens: Multi-aspect fine-grained evaluation for medical text generation. *arXiv preprint arXiv:2311.09581*, 2023.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Curtis Langlotz, and Jean-Benoit Delbrouck. Overview of the first shared task on clinical text generation: RRG24 and "Discharge Me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Shaokang Yang, Jianwei Niu, Jiyan Wu, and Xuefeng Liu. Automatic medical image report generation with multi-view and multi-modal attention mechanism. In *International Conference on Algorithms and Architectures for Parallel Processing*, pp. 687–699. Springer, 2020.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision – ECCV 2022*, volume 13696 of *LNCS*, pp. 521–539, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5. doi: 10.1007/978-3-031-20059-5_30.

Kehn E. Yapp, Patrick Brennan, and Ernest Ekpo. The effect of clinical history on diagnostic imaging interpretation–a systematic review. *Academic Radiology*, 29(2):255–266, 2022.

Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel C. Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology. *arXiv preprint arXiv:2402.14252*, 2024.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Radiology report expert evaluation (ReXVal) dataset (version 1.0.0). PhysioNet, 2023a.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802, September 2023b. doi: 10.1016/j.patter.2023.100802.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, pp. 721–729. Springer, 2019.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–198. Springer, 2023.

Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. MedRG: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*, 2024.

Listing 1: Instruction to GPT-4 for extracting single-finding sentences from narrative reports.

```
System: You are an AI radiology assistant. You are helping process reports from chest X-rays.

Please extract phrases from the radiology report which refer to objects, findings, or
anatomies visible in a chest X-ray, or the absence of such.

Rules:
- If a sentence describes multiple findings, split them up into separate sentences.
- Exclude clinical speculation or interpretation (e.g. "... highly suggestive of pneumonia").
- Exclude recommendations (e.g. "Recommend a CT").
- Exclude comments on the technical quality of the X-ray (e.g. "there are low lung volumes").
- Include mentions of change (e.g. "Pleural effusion has increased") because change is
visible when we compare two X-rays.
- If consecutive sentences are closely linked such that one sentence can't be understood
without the other one, process them together.

The objective is to extract phrases which refer to things which can be located on a chest X-
ray, or confirmed not to be present.
```

A   Extended methods

A.1   Datasets used to re-train Rad-DINO

Table A.1 shows the list of datasets used to train RAD-DINO for MAIRA-2. There is no overlap between the training, validation, or test patients between the datasets in Table A.1 and Table 1.

Table A.1: Datasets used to train RAD-DINO, our image encoder.

Data source	Num. images
BRAX (Reis et al., 2022)	41 260
ChestX-ray8 (Wang et al., 2017)	112 120
CheXpert (Irvin et al., 2019)	223 648
MIMIC-CXR (Johnson et al., 2019a)	368 960
PadChest (Bustos et al., 2020)	136 787
USMix (private)	521 608
Total	1 404 383

A.2   Extraction of sentences from reports

We convert narrative reports (specifically the *Findings* section) into lists of sentences, wherein each sentence should mention at most one finding. We do this in two places: 1) Construction of grounded reports, as described in Section 2.1, 2) To enable the use of RadFact on narrative reports, since it operates on lists of sentences.

In Listings 1 and 2 we show the system message and one of the few-shot examples used for this task. Due to space limitations, the complete set of few-shots will be shared alongside the metric implementation here: https://github.com/microsoft/RadFact.

We use GPT-4 for this task, through an Azure OpenAI endpoint.

This process changes the distribution of words and even pathologies in the reports. To quantify this, in Table A.2 we convert the MIMIC validation set to lists of sentences, and compare it to the original reports

Listing 2: One of the examples provided to GPT-4 for extracting single-finding sentences from narrative reports. Speculative statements are removed and sentences with multiple findings are split into separate sentences. Here 'orig' refers to the original sentence, and 'new' is a list of derived sentences. The processed version of the report is the concatenation of the 'new' sentences.

```
User:
The lungs remain hyperinflated, in keeping with history of asthma. Since the prior study,
there are streaky opacities in the bilateral lung bases which may be due to bronchial wall
thickening, small airways disease without discrete lobar consolidation. No pleural effusion or
 pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.


Assistant:
orig: The lungs remain hyperinflated, in keeping with history of asthma.
new:
- The lungs remain hyperinflated.
orig: Since the prior study, there are streaky opacities in the bilateral lung bases which may
 be due to bronchial wall thickening, small airways disease without discrete lobar
consolidation.
new:
- Since the prior study, there are streaky opacities in the bilateral lung bases.
orig: No pleural effusion or pneumothorax is seen.
new:
- No pleural effusion is seen.
- No pneumothorax is seen.
orig: The cardiac and mediastinal silhouettes are unremarkable.
new:
- The cardiac silhouette is unremarkable.
- The mediastinal silhouette is unremarkable.
```

Table A.2: Conversion of reports into lists of sentences alters the distribution of words and pathologies. We use typical report generation metrics to compare the modified reports with the originals, using the MIMIC-CXR validation set.

Metric	Modified
ROUGE-L	82.1 [81.8, 82.5]
RG_{ER}	91.2 [90.9, 91.5]
CheXpert, uncertain as negative:	
Macro F_1-14	87.0 [86.3, 87.7]
Macro F_1-5	93.6 [92.7, 94.3]
Recall - Atelectasis	91.2 [89.8, 92.6]
Recall - Cardiomegaly	96.2 [95.2, 97.1]
Recall - No Finding	96.9 [94.6, 96.7]
Recall - Pneumonia	3.4 [1.2, 6.7]

Listing 3: System message used for `RadFact`, instructing the LLM to assess the correctness of a single sentence given a list of reference sentences.

```
System: You are an AI radiology assistant. Your task is to assess whether a statement about a
 chest X-ray (the "hypothesis") is true or not, given a reference report about the chest X-ray
. This task is known as entailment verification. If the statement is true ("entailed")
according to the reference, provide the evidence to support it.
```

using standard report generation metrics. For pathology-level CheXbert metrics, specificity is above 97% for all finding classes, indicating the conversion into sentence lists does not produce *additional* mentions of findings. For most findings, the recall is similarly high, indicating there is little loss. The notable exception is pneumonia, where the recall is $\approx 3.4\%$, indicating that over 96% of mentions of pneumonia in the original reports have been removed by this processing. This is desired via the prompt and occurs because pneumonia is a clinical interpretation of other findings, often described with speculative language such as '... opacity suggesting pneumonia'.

A.3 RadFact metric

Listings 3 to 5 show the system message, sample few-shot examples, and a sample query for `RadFact`. The LLM is prompted to produce valid YAML outputs that can easily be parsed, which is enforced with Pydantic[12] via LangChain[13]. As in Appendix A.2, due to space limitations we show only one of the few-shot examples – the rest can be found in the repository[14]. Following chain-of-thought style prompting (Wei et al., 2022), we found that prompting the assistant to provide the evidence before the classification ("status") improved performance.

Using Llama3-70B as a backbone instead of GPT-4 – as in Chaves et al. (2024) – provides multiple advantages: It is open-source and faster, making it more accessible to the research community and advantageous when evaluating large volumes. In Table A.3, we compare the performance and throughput of `RadFact` using Llama3-70B and GPT-4. We measure performance on the binary task of entailment verification: classifying a given hypothesis sentence as entailed or not, given a list of references. In practice, to compute `RadFact` we need to process one such query per sentence in the report, in each direction. This results in, on average, six to seven queries per report. In this light, the performance drop observed in Table A.3 seems neglegible relative to the gain in throughput.

---

[12] https://github.com/pydantic/pydantic
[13] https://www.langchain.com/
[14] https://github.com/microsoft/RadFact

Listing 4: Two of the examples used in the entailment verification task in `RadFact`. The model is tasked with assigning a logical status (either `entailment` or `not_entailment`) to the hypothesis sentence, given the list of reference sentences. The 'evidence' field is a list of reference sentences supporting the logical state. For 'Degenerative changes are seen throughout the spine', nothing in the reference sentences indicates this is true, so it is labelled with `not_entailment`. 'There is persistent consolidation in the left lung base' implies 'Left basilar consolidation is present', so it is labelled with `entailment`. Note that the reverse does not hold, due to the additional detail of persistence.

```
User:
reference:
- The lungs are clear.
- The cardiomediastinal silhouette is unremarkable.
- There are no pleural effusions.

hypothesis: Degenerative changes are seen throughout the spine.

Assistant:
phrase: Degenerative changes are seen throughout the spine.
evidence: []
status: not_entailment
----------------------
User:
reference:
- There are small bilateral pleural effusions that have developed since prior study.
- Left basilar consolidation is present.
- The cardiomediastinal silhouette is within normal limits.
- No acute chest wall abnormality is radiographically evident.

hypothesis: Left basilar consolidation is present.

Assistant:
phrase: Left basilar consolidation is present.
evidence:
- There is persistent consolidation in the left lung base.
status: entailment
```

Listing 5: An example query to `RadFact`. Based on the reference sentences, the model must determine the logical state of the hypothesis.

```
User:
reference:
- The heart is borderline in size.
- There is no evidence of CHF.
- No infiltrate.
- The diaphragm is well-visualized.

hypothesis: There is a new abnormal density filling most of the right hemithorax.
```

Table A.3: Accuracy and speed of `RadFact` using Llama3-70B and GPT-4 as backbones. Llama3 runs on a single compute node with four A100 GPUs. GPT-4 is hosted on Microsoft Azure.

	Accuracy (%)	Inference speed (s/report)
Llama3	92.0	17.35
GPT-4	93.2	27.06

Table B.1: **Findings generation performance on IU-Xray.** We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Model	ROUGE-L	BLEU-4	CheXbert		RadFact Logical	
			Macro F_1-14	Micro F_1-14	Precision	Recall
MAIRA-2	27.4 [27.1, 27.8]	11.7 [11.4, 12.0]	28.6 [26.2, 31.2]	52.9 [51.3, 54.5]	71.1 [70.3, 71.9]	67.3 [66.5, 68.0]
MAIRA-2 13B	27.8 [27.4, 28.1]	11.8 [11.5, 12.1]	30.0 [27.3, 32.6]	52.5 [50.9, 54.2]	71.1 [70.3, 71.8]	68.8 [68.1, 69.5]
LLaVA-Rad	25.3 [25.0, 25.7]	−	−	53.5 [51.6, 55.8]	−	−

`RadFact-Llama3` shows high alignment with the errors spotted by radiologists in the ReXVal dataset (Yu et al., 2023a). The Kendall rank correlation coefficient between the error counts in ReXVal and the logical F1-score of `RadFact` (computed as the harmonic mean between the logical precision and the logical recall) is 0.59 [0.51, 0.66] (0.62 [0.55, 0.68] for clinically significant errors). Confidence intervals were computed using bootstrapping with $n = 1000$ in concordance with Yu et al. (2023b). While the correlation of `RadFact` is smaller than of the recently proposed CheXprompt (Chaves et al., 2024), the latter presents an attempt to directly count the different errors using a LLM. In contrast, `RadFact` is not restricted to the six error types defined in ReXVal, and can perform entailment verification for any sentence that can potentially occur in a report, naturally leading to a lower alignment with ReXVal. We found, for example, mentions of lateral images in reports from all datasets used for training MAIRA-2. Hallucinations or omissions of such mentions would not be detected by CheXprompt.

B Extended results

B.1 Findings generation on IU-Xray

Table B.1 shows external validation performance on IU-Xray. We compare to LLavA-Rad (Chaves et al., 2024) to provide a fair comparison of *held-out* performance on the *Findings* generation task. MAIRA-2 produces higher ROUGE-L scores and statistically equivalent CheXbert Micro F_1-14 scores. One risk associated with using additional inputs (such as the *Technique* and *Comparison* sections, which LLaVA-Rad does not use) is that MAIRA-2 would over-rely spurious, dataset-level associations between these inputs and the *Findings* section. However, our findings on IU-Xray suggest this has not occurred to a significant degree. In particular, the high `RadFact` scores suggest that MAIRA-2 may be producing higher-quality reports than it does on MIMIC-CXR, however this may also reflect that IU-Xray is an 'easier' dataset than MIMIC-CXR.

B.2 Impact of multitask training

Tables B.2 and B.3 provide tabular versions of the results shown in Figs. 4 and 5, indicating the impact of dropping either FindGen or GroundRep tasks from the multitask training mix.

B.3 Further ablations on additional inputs

The ablations in this section provide further insight on the individual effect of different views and sections.

Description of the '%Comparison mentions' and '%Lateral mentions' metrics   We used a language model (`Llama3-70B-Instruct`) to detect if a findings-section mentions a comparison to a prior report. We

Table B.2: Impact of dropping the `FindGen` task during training on `GR-Bench` grounded reporting performance. These tables mirror Fig. 4. The top table shows text-based metrics, while the bottom table shows box and grounding-based metrics. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Experiment	ROUGE-L	CheXbert Macro F_1-14	RG_{ER}	RadCliQ ($\downarrow$)	RadFact Logical Precision	Recall
MAIRA-2	58.2 [56.7, 59.8]	40.9 [35.9, 47.1]	56.9 [55.3, 58.5]	1.63 [1.55, 1.7]	73.5 [72.2, 74.9]	72.4 [71.0, 73.8]
NoFindGen	55.6 [53.9, 57.0]	19.6 [16.7, 23.4]	53.1 [51.5, 54.7]	1.86 [1.79, 1.93]	68.9 [67.5, 70.4]	64.9 [63.4, 66.4]

Experiment	RadFact Grounding Precision	Recall	Box-completion Precision	Recall	IoU
MAIRA-2	68.2 [64.7, 71.7]	92.2 [89.8, 94.4]	68.4 [67.2, 69.7]	84.6 [83.7, 85.5]	60.7 [59.4, 61.9]
NoFindGen	74.3 [70.2, 78.5]	92.5 [89.6, 95.1]	66.3 [64.9, 67.6]	82.7 [81.8, 83.6]	58.4 [57.1, 59.5]

Table B.3: Impact of dropping the `GroundRep` task during training on MIMIC *Findings* generation performance. This table mirrors Fig. 5. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Experiment	ROUGE-L	CheXbert Macro F_1-14	RG_{ER}	RadCliQ ($\downarrow$)	RadFact Logical Precision	Recall
MAIRA-2	38.4 [37.8, 39.1]	42.7 [40.9, 44.4]	51.5 [49.3, 53.5]	39.7 [38.9, 40.4]	2.64 [2.61, 2.68]	50.5 [49.7, 51.3]
NoGroundRep	38.3 [37.7, 38.9]	41.8 [40.2, 43.8]	49.9 [47.7, 51.7]	39.6 [39.0, 40.3]	2.65 [2.61, 2.68]	51.2 [50.4, 52.1]

evaluated the prior detection algorithm on 100 samples from the training sets of each MIMIC-CXR, PadChest, and USMix, and found it very robust with 98%, 96%, and 97% accuracy, respectively. The evaluation sets were balanced w.r.t. the prevalence of prior mentions. Since mentions of lateral images in the findings are usually explicit, we resorted to a simple regex shown in Listing 6 and refrained from creating evaluation sets for this task. Applying these algorithms to the generated and reference findings allows us to estimate in how many cases the model should have, and in how many cases it has mentioned a prior report or lateral image. Logical precision or recall values as in `RadFact` can not be computed from these numbers, as the detected mentions in the prediction and the reference do not have to be related.

Inputs containing temporal information In Table B.4, we show training and inference-time ablations demonstrating the independent effect of including the prior study and comparison section. As in Section 3.4, this analysis is performed on the subset of the MIMIC test set that has prior images. When we train without the prior study 'Train:No *Prior*', we observe a significant drop in Macro F_1-14 (-8.5%). We also see a similar but larger drop in Macro F_1-14 (-10.3%) when we train with the prior study but drop it during inference 'Infer:No *Prior*', indicating that MAIRA-2 uses the prior study to produce more factually correct reports. We also note that using a model trained with prior studies and running inference without prior studies will cause fewer hallucinations of comparisons (72.9%) as compared to a model that was not trained with prior studies (82.8%). When we train a model without the comparison section 'Train:No *Comp*', we observe a

Listing 6: Regular Expression used to detect mentions of lateral images.

```
(pa|ap|frontal) and lateral|
\blateral and (pa|ap|frontal)|
\blateral (projection|view)|
(two|2) views
```

Table B.4: Prior and comparison ablation experiments on MIMIC-CXR, on the set of test cases with a prior study (n=2181). No *Comp* means we drop the comparison section, No *Prior* means we drop the prior frontal image and the prior report. Infer:means we drop the inputs only at inference time (we evaluate a model trained using these inputs), otherwise we both train and evaluate without the inputs. This table complements Fig. 6. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Experiment	ROUGE-L	CheXbert Macro F_1-14	RadCliQ (↓)	RadFact Logical Precision	Recall	% Mentions comparison
MAIRA-2	38.4 [37.7, 39.0]	43.7 [41.9, 45.6]	2.64 [2.61, 2.68]	52.6 [51.4, 53.6]	48.6 [47.4, 49.7]	85.6 [84.2, 87.0]
Infer:No *Comp*	29.8 [29.2, 30.4]	39.9 [38.0, 41.6]	3.07 [3.03, 3.10]	47.9 [46.7, 49.0]	42.6 [41.7, 43.8]	71.2 [69.2, 73.2]
Train:No *Comp*	34.9 [34.3, 35.5]	41.9 [39.9, 43.7]	2.81 [2.78, 2.85]	52.7 [51.6, 53.7]	46.4 [45.3, 47.4]	86.4 [84.9, 87.9]
Infer:No *Prior*	37.9 [37.3, 38.6]	39.2 [37.6, 41.2]	2.69 [2.66, 2.73]	51.5 [50.5, 52.5]	47.1 [46.2, 48.2]	72.9 [71.1, 74.8]
Train:No *Prior*	38.2 [37.5, 38.9]	40.0 [38.2, 41.8]	2.67 [2.63, 2.71]	52.5 [51.6, 53.6]	47.4 [46.4, 48.4]	82.8 [81.2, 84.3]
Infer:No *Prior* No *Comp*	27.3 [26.7, 28.0]	35.8 [34.2, 37.5]	3.18 [3.15, 3.22]	45.5 [44.4, 46.5]	40.5 [39.6, 41.4]	38.6 [36.7, 40.5]
Train:No *Prior* No *Comp*	33.9 [33.2, 34.5]	39.3 [37.5, 41.1]	2.89 [2.86, 2.93]	50.6 [49.5, 51.5]	44.7 [43.7, 45.7]	75.8 [73.9, 77.4]

significant drop in lexical metrics (-9.1% drop in ROUGE-L) as well as an increase in RadCliQ (+6.4), but no significant drop in Macro F_1-14. When we train with the comparison section but drop it at inference time 'Infer:No *Comp*', we note an even larger drop in ROUGE-L (-22.4) in addition to an overall decrease in performance across all other metrics. Based on the large drop in lexical metrics when not using the comparison section, and the reduction in hallucinations when we train a model with comparison sections and run inference without them 'Infer:No *Prior* No *Comp*' as compared to training without these sections entirely 'No *Prior* No *Comp*', we hypothesise that the model uses the comparison section as an indicator of whether or not temporal change mentions should be generated in the text, and that the prior image is necessary to ensure the change words generated are correct.

Table B.5: Lateral and technique ablations on MIMIC-CXR for the subset of the test set with a lateral view (n = 1,116). No *Lat* means we drop the lateral view, No *Tech* means we drop the *Technique* section. 'Inf' means we drop the inputs only at inference time, evaluating a model trained using those inputs. Otherwise, we both train and evaluate without the inputs. This table complements Fig. 7. We report median and 95% confidence intervals based on 500 bootstrap samples. '↓' indicates that lower is better. CheXpert F_1 metrics are computed based on CheXbert labeller outputs. `RadFact` uses `RadFact-Llama3`.

Experiment	ROUGE-L	CheXbert Macro F_1-14	RadCliQ (↓)	RadFact Logical Precision	Recall	% Mentions lateral
MAIRA-2	40.4 [39.5, 41.4]	38.8 [35.9, 41.5]	2.50 [2.44, 2.56]	60.2 [58.7, 61.7]	54.7 [53.2, 56.0]	39.6 [36.6, 42.5]
Infer:No *Lat*	38.9 [38.0, 39.8]	36.8 [34.5, 39.6]	2.54 [2.49, 2.60]	60.5 [59.0, 61.9]	53.2 [51.7, 54.6]	13.2 [11.3, 15.3]
Train:No *Lat*	40.4 [39.4, 41.4]	39.1 [36.5, 42.5]	2.51 [2.46, 2.56]	60.4 [58.9, 62.0]	54.1 [52.8, 55.6]	38.2 [35.4, 41.3]
Infer:No *Tech*	34.1 [33.2, 34.9]	36.6 [33.7, 39.5]	2.81 [2.76, 2.86]	56.6 [55.2, 58.2]	51.1 [49.8, 52.3]	61.2 [58.4, 64.0]
Train:No *Tech*	37.2 [36.3, 38.2]	36.1 [33.3, 38.6]	2.64 [2.59, 2.69]	58.4 [57.0, 59.7]	53.1 [51.8, 54.4]	36.3 [33.2, 39.4]
Infer:No *Lat* No *Tech*	31.5 [30.8, 32.3]	35.2 [32.8, 37.7]	2.87 [2.82, 2.92]	57.7 [56.3, 59.0]	49.1 [47.8, 50.7]	5.1 [3.8, 6.4]
Train:No *Lat* No *Tech*	36.8 [36.0, 37.9]	39.0 [36.2, 41.9]	2.66 [2.61, 2.71]	58.5 [57.2, 59.9]	53.7 [52.3, 55.0]	36.1 [33.0, 39.2]

Inputs related to multi-view studies Table B.5 shows training and inference-time ablations to evaluate the impact of including the lateral view and the technique section independently, as a complement to the analysis in Section 3.4 demonstrating their joint effect. Similarly, we restrict the analysis to the test studies that include a lateral view (n=1,116, 30.6%). When we drop the lateral view at inference-time 'Infer:No *Lat*', we notice that MAIRA-2 generates less lateral mentions (13.23% vs 39.57%) and therefore limited "lateral hallucinations". We also observe a drop in almost all metrics including Macro F_1-14 (-5.15%) highlighting the importance of the lateral view in making accurate diagnosis. On the other hand, a model trained without the lateral view continues to hallucinate lateral mentions (38.16%) since it can use the technique section as a proxy to make simple lateral predictions. Even though this ablated model is able to generate simple lateral

references using the technique section as a shortcut, there is no guarantee that it has improved it's clinical accuracy when a pathology can only be seen on the lateral. Moreover, when we drop the technique in the presence of the lateral view 'Infer:No *Tech*', we see a large drop in ROUGE-L (-15.59%) and a substantial increase of the *%Lateral mentions*, exceeding 35.57% (percentage of lateral mentions in the ground truth) by a very large margin. This suggests that the technique section is a strong indicator for generating lateral mentions. However, when this information is omitted during training 'Train:No *Tech*', the model can still figure out when to mention the lateral view (36.33%) but not as accurately as in MAIRA-2. Finally, when we drop both the lateral and the technique at the same time during inference 'Infer:No *Lat* No *Tech*', the percentage of lateral mentions drops down to 5.1% (getting closer to 0) indicating that both the lateral view and the technique section are essential to reduce hallucinations related to lateral mentions. This further becomes clearer when compared to a model that is trained without this information 'Train:No *Lat* No *Tech*' but still hallucinates lateral mentions (36.12%) as discussed in Section 3.4.

C Additional qualitative examples

C.1 Successful grounded reporting examples from GR-Bench

We showcase additional sample generations from MAIRA-2 with comments from radiologist review in Figures C.1 to C.3.

C.2 High and low-scoring examples from GR-Bench according to RadFact

Figures C.4 to C.6 present manually selected qualitative example of MAIRA-2 output on GR-Bench with varying RadFact logical precision: 1.0, 0.78 and 0.0 respectively. Figures C.7 to C.9 present additional examples selected based on varying RadFact grounding precision: 1.0, 0.5 and 0.0 respectively.

C.3 Findings generation examples from MIMIC-CXR

It is not possible to quantitatively compare to models trained to generate other sections, such as *Impression* (Bannur et al., 2023) or both *Findings* and *Impression* together (Tanno et al., 2023; Yang et al., 2024). In Figures C.10 to C.13, we qualitatively compare on the examples shown in Yang et al. (2024), which were sourced from the MIMIC-CXR validation set. We find that all four study examples represent mostly "normal" patient cases that make little or no references to prior or lateral images. As illustrated in the model outputs, there's little difference between MAIRA-2 and Med-Gemini phrases, and the original reference text. In independent reviews with two radiologists, minor variances were surfaced in terms of findings missed or hallucinated, and preferences for their conciseness or ordering that are described in the Figure captions. Overall, for this very limited set of examples, which predominantly report negative rather than more clinically relevant (positive) findings, it is difficult to surface any more clinically significant differences between the outputs of either model.

Figure C.1: A manually selected qualitative example of MAIRA-2 output on GR-Bench. This 3-part figure shows MAIRA-2 model inputs (top); the MAIRA-2 phrase outputs vis-à-vis the reference text (middle); and grounding boxes for the MAIRA-2 phrases on the current frontal image alongside their alignment with reference boxes (bottom). The selected example has moderate MAIRA-2 RadFact logical precision (0.78) and recall (0.75). Qualitative comparison with the reference text suggests that MAIRA-2 misclassified the patient's bony structures as "intact"; added that the uncoiling of the aorta is "compatible with hypertension"; and missed detecting the "degenerative changes of the thoracic spine" and that the "lungs are hyperlucent". In individual reviews with two consultant radiologists, it was suggested that the demineralisation of the bony structures is difficult to see on the images and therefore considered a borderline finding to call out. Similarly, the degenerative changes of the spine were assessed as only mild. Furthermore, the addition of hypertension was regarded as 'acceptable' since the aorta is slightly torturous. Lastly, it was noted how the MAIRA-2 findings also included that "There is a pectus deformity of the sternum", which was not reported in the reference and can only be clearly seen on the lateral view. For image grounding, there was no overlap between four abnormal findings that were reported in either the MAIRA-2 candidate or the reference text, resulting in non-corresponding bounding box as is reflected in lower spatial precision (0.67) and recall (0.4) scores. For the two abnormal findings that were reported and entailed in both findings texts, however, there is high grounding precision and recall (1.0).

MAIRA-2 model inputs:

Frontal   Lateral (flipped horizontally for display)

INDICATION: R09.89 Abnormal Chest Sounds R63.0 Anorexia.

TECHNIQUE: PA and lateral views of the chest were obtained.

COMPARISON: None.

MAIRA-2 model outputs:

RadFact logical precision: 0.75
RadFact logical recall: 0.71

FINDINGS phrases

A. There is a moderatesized right pleural effusion.
B. Small amount of fluid is seen in the minor fissure.
C. There is likely consolidation or atelectasis present in the right middle lobe and right lung base.
D. No vascular congestive changes.
E. Median sternotomy sutures noted.
F. Prosthetic intracardiac valve noted.
G. There is no hilar or mediastinal adenopathy.
H. The bony structures are intact.

Reference text

1. Dense consolidation seen in the right middle lobe.
2. Blunting of right lateral costophrenic sulcus.
3. The cardiac size is normal.
4. Mediastinal sutures noted.
5. Prosthetic mitral valve noted.
6. There is no hilar or mediastinal adenopathy.
7. The bony structures are intact.

Grounded FINDINGS

Box alignment

RadFact spatial precision: 0.33
RadFact spatial recall: 0.5
RadFact grounding precision: 0.5
RadFact grounding recall: 1.0

FINDING not entailed in either reference or predicted phrases

B. Small amount of fluid is seen in the minor fissure.

2. Blunting of right lateral costophrenic sulcus.

Text variances
MAIRA-2 box
Reference box

FINDINGS entailed (uni-directionally)

A. There is a moderatesized right pleural effusion. <-- 2. Blunting of right lateral costophrenic sulcus.

C. There is likely consolidation or atelectasis present in the right middle lobe and right lung base. <-- 1. Dense consolidation seen in the right middle lobe. 2. Blunting of right lateral costophrenic sulcus.

1. Dense consolidation seen in the right middle lobe. <-- C. There is likely consolidation or atelectasis present in the right middle lobe and right lung base.

Figure C.2: A manually selected qualitative example of MAIRA-2 output on GR-Bench. This 3-part figure shows MAIRA-2 model inputs (top); the MAIRA-2 phrase outputs vis-à-vis the reference text (middle); and grounding boxes for the MAIRA-2 phrases on the current frontal image alongside their alignment with reference boxes (bottom). The selected example has moderate MAIRA-2 RadFact logical precision (0.75) and recall (0.71). In this example study, MAIRA-2 model outputs state moderate right pleural effusion, small amount of fluid in the minor fissure; as well as the presence of consolidation or atelectasis in the right middle lobe and right lung base. In review with a consultant radiologist, they agreed with these findings, however, they found that the corresponding MAIRA-2 bounding boxes for findings B and C were too big. For example, a small amount of fluid in the minor fissure is only visible as a small single line in the middle of the much larger box for finding C. As such, this study presents an example of good logical precision, however, with lower spatial performance. Both the reference text and the MAIRA-2 outputs also state different normals (e.g., normal cardiac size, no vascular congestive changes). Furthermore, reviewing finding E and reference finding 4, the consultant radiologist preferred the MAIRA-2 phrase of "Median sternotomy sutures noted.", since it is more accurate in its indication of the sutures: sternotomy rather than the mediastinal. Regarding finding 5, the term "mitral" simply presents a type of "intracardiac valve", and therefore finding F was considered acceptable.

Figure C.3: A manually selected qualitative example of MAIRA-2 output on GR-Bench. This 3-part figure shows MAIRA-2 model inputs (top); the MAIRA-2 phrase outputs vis-à-vis the reference text (middle); and grounding boxes for the MAIRA-2 phrases on the current frontal image alongside their alignment with reference boxes (bottom). The selected example has moderate MAIRA-2 RadFact logical precision (0.75) and recall (0.86). Both the reference text and MAIRA-2 phrase output suggest the existence of a small left effusion, which can be clearly seen on the lateral view. On the frontal image, the irregularity of the diaphragm suggests that there is small infiltrate at the left base. The identified infiltrate and effusion are considered to explain well the symptoms of chest pain and cough that are given in the indication; and the grounding box for finding B is evaluated to be appropriate for the finding. Nonetheless, MAIRA-2 findings erroneously state that the diaphragm is smooth, when it has irregularities. Whilst not mentioned in the reference text, MAIRA-2 outputs also include "The pulmonary interstitial markings are normal.", which is correct. In this instance, the reference boxes for findings 2 and 3, which were drawn by human annotators, are very small. Consequently, even though there was good logical entailment for the key abnormal findings, their corresponding boxes did not overlap enough (given the set 50% threshold), explaining the low grounding precision scores.

MAIRA-2 model inputs:

Frontal   Lateral

MAIRA-2 model outputs:

RadFact logical precision: 1.0
RadFact logical recall: 1.0

FINDINGS phrases

A. The heart is within normal limits.
B. The lungs are within normal limits.
C. The mediastinal structures are within normal limits.
D. There is mild uncoiling of the aorta unchanged.
E. There are no osseous abnormalities.

Reference text

1. The heart is normal in size.
2. Uncoiled aorta.
3. The lungs are well-inflated.
4. There is no chf.
5. There is no infiltrate.
6. There is no mass lesion.
7. There are no osseous abnormalities.

▇ Highlighted text variances
☐ MAIRA-2 box

Figure C.4: This example has high `RadFact` logical precision and recall (1.0) and presents an interesting case since the patient has "situs inversus", meaning all organs are mirrored in the body (e.g., the heart, aorta and stomach are on the right rather than the left side). Neither the reference text nor MAIRA-2 outputs state "situs inversus", an observation that is, of course, within normal limits. The study is mostly normal and findings well entailed. However, while the study information state no comparison, the MAIRA-2 output hallucinated "unchanged" about the uncoiled aorta. The reference text, relating to a normal study, also does not have any box annotations, meaning that the uncoiled aorta is only grounded within the MAIRA-2 findings.

MAIRA-2 model inputs:

Frontal

Lateral (flipped horizontally for display)

INDICATION: Z01.812 Encounter for preprocedural laboratory examination.

TECHNIQUE: PA and lateral views of the chest were obtained.

COMPARISON: None.

MAIRA-2 model outputs:

RadFact logical precision: 0.78
RadFact logical recall: 0.64

FINDINGS phrases

A. The lungs show no active infiltrate.
B. The lungs show no mass.
C. The lungs show no effusion.
D. Diaphragms are sharp.
E. Minor fibrotic/subsegmental atelectatic changes at the left base are noted.
F. The cardiac size is mildly enlarged.
G. There is no hilar adenopathy.
H. There is no mediastinal adenopathy.
I. The bony structures are intact.

■ Highlighted text variances
☐ MAIRA-2 box
⌐⌐ Reference box

Reference text

1. The lungs show no active infiltrate.
2. The lungs show no mass.
3. The lungs show no effusion.
4. Diaphragms are sharp.
5. The heart is slightly enlarged.
6. Atherosclerotic calcifications of the aorta are present.
7. There is no hilar adenopathy.
8. There is no mediastinal adenopathy.
9. Healed fracture of the left clavicle is present.
10. Healed fractures of the left upper ribs are noted.
11. There is a bullet projecting in the inferior right lower chest.

Figure C.5: This example has moderate RadFact logic precision (0.78) and recall (0.64). Many of the phrases are well-matched. MAIRA-2 output however missed the "bullet" that is projecting in the right lower chest, and it does not include the calcification of the aorta, which was described in review with a consultant radiologist as a very difficult to see finding and therefore a borderline observation. Where the reference states healed fractures, MAIRA-2 outputted that the bony structures are intact. MAIRA-2 outputs further include "Minor fibrotic/subsegmental atelectatic changes at the left base are noted"; which is evidenced by the elevated left hemidiaphragm pushing in the lung with resulting atelectasis – a finding that was not reported in the reference text.

MAIRA-2 model inputs:

Frontal   Lateral

INDICATION: COPD PULMONARY INFILTRATE.

TECHNIQUE: CHEST TWO VIEW.

COMPARISON: None.

MAIRA-2 model outputs:

RadFact logical precision: 0.0
RadFact logical recall: 0.0

FINDINGS phrases

A. The heart is stable.
B. Pulmonary vascularity is unremarkable.
C. No infiltrate is seen.
D. No pleural effusion is seen.

Reference text

1. Minimal scarring or subsegmental atelectasis is noted at the left lung base.
2. Lungs remain mildly hyperinflated.

□ Highlighted text variances
⌐⌐ Reference box

Figure C.6: This example has low RadFact logical precision and recall (0.0). MAIRA-2 phrases suggest this is a normal study, whilst the reference states "Minimal scarring or subsegmental atelectasis is noted at the left lung base." and "Lungs remain mildly hyperinflated.". Both present minimal or mild findings that were however missed. In reviews with a consultant radiologists it was pointed out that the study indication states COPD, which – where it is a known condition – would mean hyperinflation is to be expected. Furthermore, the review surfaced that both text candidates missed the "scoliosis" – a sideways curvature of the spine – that is visible in the frontal image.

MAIRA-2 model inputs:

Frontal

Lateral (flipped horizontally for display)

INDICATION: I10 Hypertension E11.9 type 2 diabetes.

TECHNIQUE: PA and lateral views of the chest were obtained.

COMPARISON: None.

MAIRA-2 model outputs:

RadFact logical precision: 0.9
RadFact logical recall: 1.0

FINDINGS phrases

A. The bony structures are intact.
B. There are degenerative changes of the spine.
C. The lungs show no acute infiltrate or mass.
D. There is no effusion.
E. The pulmonary interstitial markings are normal.
F. The diaphragm is smooth.
G. The cardiac size is mildly enlarged.
H. There is uncoiling of the aorta.
I. Calcification is noted within the aortic knob.
J. There is no hilar or mediastinal adenopathy.

Reference text

1. The bony structures are intact.
2. There are degenerative changes of the spine.
3. The lungs show no acute infiltrate or mass.
4. There is no effusion.
5. The pulmonary interstitial markings are normal.
6. The diaphragm is smooth.
7. The cardiac size is mildly enlarged.
8. Considerable uncoiling the aorta is noted.
9. There is no hilar or mediastinal adenopathy.

Grounded FINDINGS

Box alignment

RadFact spatial precision: 0.75
RadFact spatial box recall: 1.0
RadFact grounding precision: 1.0
RadFact grounding box recall: 1.0

FINDING not entailed in reference phrases

I. Calcification is noted within the aortic knob. I. Calcification is noted within the aortic knob.

Text variances
MAIRA-2 box
Reference box

FINDINGS entailed (bi-directionally)

B./ 2. There are degenerative changes of the spine.

G./ 7. The cardiac size is mildly enlarged.

H. There is uncoiling of the aorta.H. There is uncoiling of the aorta. / 8. Considerable uncoiling the aorta is noted.

Figure C.7: This example has high grounding precision (1.0). There is generally high overlap between both findings texts. MAIRA-2 output includes the finding of a "Calcification is noted within the aortic knob.", which is described in radiologist review as a plausible, borderline findings that was however not included in the reference text. Where generated MAIRA-2 findings and boxes are matching the reference, resulting grounding precision is high.

MAIRA-2 model inputs:

Frontal

Lateral (flipped horizontally for display)

INDICATION: 287.891 smoker. J93.9 previous pneumothorax.

COMPARISON: None.

MAIRA-2 model outputs:

RadFact logical precision: 0.91
RadFact logical recall: 0.78

FINDINGS phrases

A. The bony structures are intact.
B. The lungs show no active infiltrate.
C. The lungs show no mass.
D. The lungs show no effusion.
E. There is evidence of lung hyperexpansion.
F. There is evidence of apical pleural thickening on the right.
G. The previously reported pneumothorax has cleared.
H. The diaphragm is smooth.
I. The cardiac size is normal.
J. There is no hilar adenopathy.
K. There is no mediastinal adenopathy.

Reference text

1. The bony structures are intact.
2. There is a mild scoliosis.
3. The lungs show no acute infiltrate or mass.
4. There is no effusion.
5. The pulmonary interstitial markings are normal.
6. There is right apical pleural thickening with surgical clips in the right upper lobe medially compatible with pneumothorax repair.
7. The diaphragm is smooth.
8. The cardiac size is normal.
9. There is no hilar or mediastinal adenopathy.

Grounded FINDINGS

Box alignment

RadFact spatial precision: 0.33
RadFact spatial recall: 0.5
RadFact grounding precision: 0.5
RadFact grounding recall: 1.0

FINDINGS not entailed in predicted or reference phrases

E. There is evidence of lung hyperexpansion.

2. There is a mild scoliosis.

Text variances
MAIRA-2 box
Reference box

FINDINGS entailed (uni-directionally)

F. There is evidence of apical pleural thickening on the right. <-- 6. There is right apical pleural thickening with surgical clips in the right upper lobe medially compatible with pneumothorax repair.

G. The previously reported pneumothorax has cleared. <-- 6. There is right apical pleural thickening with surgical clips in the right upper lobe medially compatible with pneumothorax repair.

6. There is right apical pleural thickening with surgical clips in the right upper lobe medially compatible with pneumothorax repair. <-- F. There is evidence of apical pleural thickening on the right. G. The previously reported pneumothorax has cleared.

Figure C.8: This example has moderate grounding precision (0.5). The MAIRA-2 outputs include a finding stating "There is evidence of lung hyperexpansion", which in radiologist review was verified to be correct. Both findings texts correctly identified the apical pleural thickening on the right upper lobe. However, the reference text expands this finding to also include commentary about surgical clips and pneumothorax repair, whereas the MAIRA-2 model outputs a separate phrase stating the "previously noted pneumothorax has cleared". Although there is good alignment with the reference boxes where the findings specify the apical pleural thickening (findings F and 6), MAIRA-2 falsely generated a whole right lung box for a cleared pneumothorax (finding G), which would not match to the much narrower reference bounding box that is centered on the pleural thickening; thereby explaining the lower entailed box performance metrics. Nonetheless, it is interesting to point out that change information about the pneumothorax was generated even though no prior image was available to this study, likely as a consequence of the study indication that states "previous pneumothorax".

Figure C.9: This example has low grounding precision (0.0). The reference text states that the finding of "linear scarring versus subsegmental atelectasis" exists bilaterally, whereas the MAIRA-2 outputs describe these as right-sided only. In radiologist review, the left lower lobe omission was indeed considered a missed subtle finding. MAIRA-2 outputs also did not include the report of the "calcification of aortic knob". In terms of bounding box placement and size, the MAIRA-2 box for finding A was considered a bit too big, but acceptable; whereas the corresponding bounding boxes for the reference text (Finding 1) were described as a bit too narrow; especially on the right side. This low overlap in bounding boxes explains the lower box precision scores in this instance.

Current Study

Frontal

No Lateral, No Prior

Reference: Original report

INDICATION: History: _F with intubation for ICH // position of ET tube

TECHNIQUE: Portable upright chest radiograph

COMPARISON: -

FINDINGS: Tip of the endotracheal tube projects over the mid thoracic trachea, approximately 3.7 cm from the carina. Enteric tube terminates beyond the diaphragm, in the left upper quadrant. Lungs are clear and cardiomediastinal silhouette is normal

MAIRA-2 outputs:

FINDINGS: Endotracheal tube terminates 4.5 cm from the carina. Enteric tube terminates in the left upper quadrant. Lungs are clear. Cardiomediastinal and hilar contours are normal. No pleural effusion or pneumothorax.

MedGemini outputs:

FINDINGS: The endotracheal tube terminates 4.5 cm above the carina. an enteric tube courses below the diaphragm and terminates within the stomach. There is no focal consolidation, pleural effusion, pneumothorax, or pulmonary edema. The cardiomediastinal silhouette is within normal limits.

Figure C.10: One of the examples from Yang et al. (2024) to enable qualitative comparison to Med-Gemini. Apart from the specific lines and tubes findings, this study is mostly normal. Neither generated report findings missed any observations and both were assessed in radiologist reviews as equivalent from a clinical perspective. Interestingly, both candidate texts hallucinated the 4.5 cm measure of the endotracheal tube terminating above the carina, as neither model can plausibly predict the correct measurement from a chest X-ray as they have no information on scale. While the MAIRA-2 output produces the exact same location description for the enteric tube as the reference text stating it to terminate "in the left upper quadrant"; the Med-Gemini location of "terminates within the stomach" is considered as more precise. On the other hand, preferences were expressed for MAIRA-2 stating "Lungs are clear", which is more concise than its counterpart; and for MAIRA-2 's ordering of the findings from lungs to the cardiomedistinal/ hilar structures and then the pleura (similar to a structured report generation), compared to the Med-Gemini findings that move between those structures.

Prior Study

Frontal

Reference: Original report

INDICATION: Patient with fever. Assess for pneumonia.

COMPARISON: _.

FINDINGS: Frontal and lateral views of the chest demonstrate normal lung volumes. Right lung base opacities are slightly more conspicuous since prior, projecting over the spine on the lateral view. There is no pleural effusion, focal consolidation or pneumothorax. Hilar and mediastinal silhouettes are unchanged. Heart size is normal. There is no pulmonary edema. Partially imaged upper abdominal visceral organs are unremarkable.

IMPRESSION: Right lung base opacities, slightly more conspicuous since _, may represent atelectasis or infection in the appropriate clinical setting.

Current Study

Frontal   Lateral

Reference: Original report

INDICATION: n/a

TECHNIQUE: Chest frontal and lateral views.

COMPARISON: "_."

FINDINGS: Frontal and lateral views of the chest were obtained. No focal consolidation, pleural effusion or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable. No pulmonary edema is seen.

MAIRA-2 outputs:

FINDINGS: Frontal and lateral views of the chest were obtained. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Cardiac and mediastinal silhouettes are unremarkable.

MedGemini outputs:

FINDINGS: The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.

Figure C.11: One of the examples from Yang et al. (2024) to enable qualitative comparison to Med-Gemini. This example study reports a "normal". Both generated candidates are near identical and match the reference text findings. Like the reference text, MAIRA-2 outputs the phrase "Frontal and lateral views of the chest were obtained." Whilst learned from such input instances, technically, this information does not present an image finding and it is already included in the *Technique* description. Neither the reference text, nor MAIRA-2 and Med-Gemini phrases include any comparison information with the prior study.

Prior Study

Frontal

Reference: Original report

INDICATION: Cough.

COMPARISON: None.

FINDINGS: There is moderate enlargement of the cardiac silhouette. The aorta is mildly tortuous and calcified. Pulmonary vascularity is not engorged. Ill-defined opacity is noted within the right lung base, which is concerning for an infectious process. There is no large pleural effusion or pneumothorax. Mild degenerative changes are noted in the thoracic spine. Multiple clips are seen within the upper abdomen.

IMPRESSION: Ill-defined opacity within the right lung base which is concerning for pneumonia. Followup radiographs after treatment are recommended to ensure resolution of this finding.

Current Study

Frontal   Lateral

Reference: Original report

INDICATION: ,_-year-old woman with renal failure, evaluate for edema or infiltrate.

TECHNIQUE: -

COMPARISON: "_."

FINDINGS: PA and lateral chest radiographs were obtained. There is an ill-defined opacity in the right lower lobe that does not obscure the right heart border. A right-sided pleural effusion is small. There is no pneumothorax. Cardiomegaly is mild. Aortic calcifications are minimal.

MAIRA-2 outputs:

FINDINGS: PA and lateral views of the chest were obtained. Cardiomediastinal silhouette including mild cardiomegaly is unchanged. There is increased pulmonary vascular congestion and mild-to-moderate pulmonary edema. A small right pleural effusion is present. There is no focal consolidation or pneumothorax.

MedGemini outputs:

FINDINGS: There is a right lower lobe consolidation. there is also a small right pleural effusion. The heart is mildly enlarged. There is mild pulmonary vascular congestion. there is no pneumothorax.

Figure C.12: One of the examples from Yang et al. (2024) to enable qualitative comparison to Med-Gemini. In this example study, like Figure C.11, MAIRA-2 outputs the *Technique*, which is in keeping with the reference report. The MAIRA-2 outputs also correctly describe the "unchanged" mild cardiomegaly. However, neither candidate findings say anything about the progression of the previously reported right lower lobe opacity, nor does the reference text explicitly describe any changes from the prior. Both generated report candidates state the existence of "pulmonary vascular congestion", which is most apparent via comparison with the prior study, and yet this finding was not included in the reference text. In the reference report, it is implied that the ill-defined opacity in the right lower lobe is a consolidation. Whilst the Med-Gemini findings indeed include "There is a right lower lobe consolidation", the MAIRA-2 outputs falsely state "There is no focal consolidation". Lastly, neither generated findings texts report the minimal "aortic calcifications", which – as a chronic finding – was reported previously, and thus, our radiologists did not consider this as a significant omission.

Prior Study

Frontals

Reference: Original report

PREAMBLE: CHEST RADIOGRAPH

INDICATION: Status post nephrectomy, evaluation for fevers. Questionable pneumonia.

COMPARISON: _.

FINDINGS: As compared to the previous examination, there is no relevant change. Minimal basal areas of atelectasis. Normal size of the cardiac silhouette without pulmonary edema. No interval appearance of new parenchymal opacities suggestive of pneumonia. The contours of the hilar and mediastinal structures are unchanged and normal.

Current Study

Frontal   Lateral

Reference: Original report

INDICATION: n/a

TECHNIQUE: n/a

COMPARISON: "_."

FINDINGS: Frontal and lateral views of the chest were obtained. The lungs are hyperinflated, with flattening of the diaphragms, consistent with chronic obstructive pulmonary disease. There is mild left basilar atelectasis. No focal consolidation is seen. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are unremarkable. Degenerative changes are seen along the spine.

MAIRA-2 outputs:

FINDINGS: PA and lateral views of the chest were obtained. The lungs are hyperinflated with flattened diaphragms suggesting underlying COPD. A stable area of scarring is noted in the lower lungs. There is no evidence of pneumonia or CHF. No pleural effusion or pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact.

MedGemini outputs:

FINDINGS: The lungs are hyperinflated with flattening of the diaphragms and increased ap diameter, suggesting chronic obstructive pulmonary disease. No focal consolidation, pleural effusion, or pneumothorax is detected. Heart and mediastinal contours are within normal limits with extensive aortic calcification.

Figure C.13: One of the examples from Yang et al. (2024) to enable qualitative comparison to Med-Gemini. Again, MAIRA-2 outputs technical details of image views as part of the *Findings* as is reflective of the reference text. Both candidate reports include the suggestion of an underlying "COPD", which presents a clinical diagnosis rather than an image finding. Med-Gemini outputs further state "increased ap diameter". Whilst this finding is not false, it likely presents a hallucination since the AP dimension can only be seen on the lateral view, which was not part of the Med-Gemini model training. The MAIRA-2 findings of "stable area of scarring is noted in the lower lungs" relates to the mild left basilar atelectasis in the reference text – a finding that was not reported by Med-Gemini. While reporting of the area of scarring and its progression from the prior ("stable") are correct in the MAIRA-2 outputs, its location description is imprecise and should state in which lower lung (singular, left) it is present. The reference text further states "Degenerative changes are seen along the spine". MAIRA-2 outputs instead state that the "Bony structures are intact". There is no commentary made about the bones in the Med-Gemini output, which – similar to MAIRA-2 – may suggest an assumed normal. In general, degenerative changes to the spine, especially with the existence of prior studies, are not considered a new finding and are therefore less important to mention. Lastly, our radiologists could not see the "extensive aortic calcification" that was described in the Med-Gemini findings and that were also not remarked on by the reference text. *Please note, for MAIRA-2, only the upper frontal image of the prior study was included into the analysis.*