MAIRA-1: A SPECIALISED LARGE MULTIMODAL MODEL FOR RADIOLOGY REPORT GENERATION

TECHNICAL REPORT

Stephanie L. Hyland*[1], Shruthi Bannur[1], Kenza Bouzid[1], Daniel C. Castro[1], Mercy Ranjit[2], Anton Schwaighofer[1], Fernando Pérez-García[1], Valentina Salvatelli[1], Shaury Srivastav[2], Anja Thieme[1], Noel Codella[3], Matthew P. Lungren[4], Maria Teodora Wetscherek[1], Ozan Oktay[1], and Javier Alvarez-Valle*[1]

[1]Health Futures, Microsoft Research
[2]Microsoft Research India
[3]Microsoft Azure AI
[4]Microsoft Health and Life Sciences

ABSTRACT

We present a radiology-specific multimodal model for the task for generating radiological reports from chest X-rays (CXRs). Our work builds on the idea that large language models (LLMs) can be equipped with multimodal capabilities through alignment with pre-trained vision encoders. On natural images, this has been shown to allow multimodal models to gain image understanding and description capabilities. Our proposed model (MAIRA-1) leverages a CXR-specific image encoder in conjunction with a fine-tuned LLM based on Vicuna-7B, and text-based data augmentation, to produce reports with state-of-the-art quality. In particular, MAIRA-1 significantly improves on the radiologist-aligned RadCliQ metric and across all lexical metrics considered. Manual review of model outputs demonstrates promising fluency and accuracy of generated reports while uncovering failure modes not captured by existing evaluation practices. More information and resources can be found on the project website: https://aka.ms/maira.

1 Introduction

Large-scale pretraining of general-purpose image and language models has enabled the development of data-efficient large multimodal models. A common paradigm is to adapt a vision encoder to a pretrained LLM with varying levels of integration, as done by LLaVA [Liu et al., 2023b], InstructBLIP [Dai et al., 2023] and Flamingo [Alayrac et al., 2022]. Here, we explore this paradigm for the specialised goal of generating radiology reports from images, namely generating the *Findings* section of a report based on a frontal chest X-ray and given the *Indication* for the study.

The *Findings* section contains the radiologist's key observations about the image. To assist in their interpretation, a radiologist may refer to prior scans from the patient, other imaging modalities or views of the chest, or the patient's clinical and medical history if available and not provided in the *Indication*. Hence, reporting on a chest X-ray requires synthesis of various data sources including multiple images taken over a period of time. For this study, we focus on the simplified single-image setting, acknowledging that the resulting model will be susceptible to generating 'hallucinated' references to prior studies [Ramesh et al., 2022, Bannur et al., 2023a] or lateral views.

Although the *Findings* section aims describe the image, the generation task crucially differs from image captioning. Chest X-ray reports must also establish the *absence* of findings, for example confirming that the insertion of a central venous line did not cause a pneumothorax. Secondly, the findings in a chest X-ray are not typical 'objects'. Radiographic findings can be subtle variations in opacity against an otherwise-normal background of overlapping structures, requiring the extraction of fine-grained details from the image. Hence, findings generation remains a challenging multimodal task, requiring both the extraction of fine-grained details from the image and the generation of nuanced, radiology-specific

*Corresponding authors: sthyland@microsoft.com, jaalvare@microsoft.com

language. A model which could generate the first draft of a radiology report has the potential to improve and expedite radiology reporting workflows [Huang et al., 2023], if it can first demonstrate a sufficient standard of clinical accuracy.

General-domain models demonstrably fail at the task of findings generation [Tu et al., 2023]. We therefore develop a *radiology-specific* large multimodal model, which we call MAIRA-1. MAIRA-1 benefits from a pre-trained language model (Vicuna-7B [Chiang et al., 2023]), a radiology-specific image encoder (RAD-DINO [Pérez-García et al., 2024]), and the use of GPT-3.5 for data augmentation. By fine-tuning this model on a publicly available dataset of chest X-rays with corresponding reports (MIMIC-CXR [Johnson et al., 2019a,b]), we demonstrate the potential of large multimodal models for specialised radiology use-cases.

Concretely, we share the following outcomes:

1. Performance competitive with existing state-of-the-art is possible without extremely large models or datasets.
2. Design choices make a difference: keys to the success of this model are the use of a domain-specific image encoder with a larger image resolution and a more complex adapter layer, as well as the use of data augmentation and the *Indications* section of the report.
3. Evaluation for this task remains challenging with heterogeneity across prior work. We report a wide variety of metrics to enable comparison, and demonstrate through stratified analyses the sensitivity of performance measures to test set characteristics.

We share examples of *Findings* sections generated by MAIRA-1 to highlight both its successes and limitations. By re-purposing pre-trained LLMs we have demonstrated what is possible in a constrained setting with a limited dataset. Relative to a radiologist, MAIRA-1 receives only a limited view of the patient, relying on a single chest X-ray and the *Indication* for the study. Hence, MAIRA-1 is a step *towards* realistic report-drafting systems, which will incorporate richer inputs such as previous images or other clinical information. We anticipate that with larger and cleaner training datasets, this flexible LLM-based approach can yield further gains.

2 Related work

Generalist foundation models The paradigm of aligning image encoders to existing LLMs has achieved great success in the general domain. An example is LLaVA [Liu et al., 2023b], which combines a vision encoder with LLM using a simple adaptation layer applied to the image embedding, and employing visual instruction tuning to achieve general-purpose visual and language understanding. InstructBLIP [Dai et al., 2023] performs similar visual instruction tuning, but leverages a more powerful Q-former and a frozen LLM. Flamingo [Alayrac et al., 2022] couples the image and text more tightly, allowing a frozen LLM to cross attend to images. Flamingo additionally benefits from training on interleaved visual-text-video data. PaLM-E [Driess et al., 2023] is a multimodal decision making model, trained on text, image and sensor data, capable of embodied reasoning and decision making.

Medical/radiology adaptation of LLMs/MLLMs There have been multiple efforts to specialise generalist foundation models to the medical domain or specifically for radiology applications. For example, Med-Flamingo [Moor et al., 2023] is based on OpenFlamingo [Awadalla et al., 2023, Alayrac et al., 2022] and was trained on images and captions from medical textbooks to perform few-shot visual question answering (VQA). Med-PaLM M [Tu et al., 2023] fine-tuned the proprietary PaLM-E model [Driess et al., 2023] on a broad collection of biomedical datasets, to build a system addressing diverse uni- and multimodal tasks over text, images, and genomics. LLaVA-Med [Li et al., 2023] proposed to adapt LLaVA [Liu et al., 2023b] with a curriculum of plain image–text pairs and generated multimodal instructions based on PubMed data. ELIXR [Xu et al., 2023] aligns a CXR encoder model, SupCon [Sellergren et al., 2022], with PaLM 2-S [Anil et al., 2023] through a multi-stage training process. The resulting model can perform tasks such as classification, semantic search, question answering and quality assurance. Radiology-GPT [Liu et al., 2023c] is a text-only model based on the Alpaca instruction-tuning framework [Taori et al., 2023], trained with radiology reports from the MIMIC-CXR dataset to perform findings-to-impression generation.

Radiology report generation Given the long tail of possible observations in a chest X-ray, and the need for fine-grained description of findings, the generation of the narrative radiology report itself from the image is a promising target for machine learning systems [Wang et al., 2018]. Such systems necessarily require a generative language modelling component; initially recurrent neural networks [Wang et al., 2018, Liu et al., 2019] giving way more recently to transformer architectures [Miura et al., 2021, Chen et al., 2020, Bannur et al., 2023a], including LLMs such as PaLM [Tu et al., 2023] and here Vicuna-7B [Chiang et al., 2023].

Given the need for clinical accuracy in the generated text, a line of work departs from a vanilla language-modelling loss and uses reinforcement learning (RL) to optimise for 'clinically relevant' rewards, based on the presence of specific

findings [Liu et al., 2019, Irvin et al., 2019], or logical consistency [Miura et al., 2021, Delbrouck et al., 2022]. A downside of such approaches is the reliance on models such as CheXbert [Smit et al., 2020] or RadGraph [Jain et al., 2021] to extract clinical entities, and the more complex optimisation problem. Here we demonstrate what is possible through plain auto-regressive language modelling alone, acknowledging that gains from more sophisticated training objectives or RL-based approaches are likely complementary.

Prior work has focused on generating different sections of the radiology report, sometimes the combination of both *Findings* and *Impression* [Wang et al., 2018, Chen et al., 2020, Jin et al., 2023, Yan et al., 2023], the *Impression* alone [Endo et al., 2021, Bannur et al., 2023a], or—as here—the *Findings* only [Miura et al., 2021, Delbrouck et al., 2022, Tanida et al., 2023, Nicolson et al., 2023, Tu et al., 2023]. Jeong et al. [2023] and Yu et al. [2023] studied all three settings, providing evidence that the choice of section markedly impacts reported metrics, prohibiting comparison between variations of the task.

3 Method

In this section, we explain details of our task definition, dataset composition and preparation, modelling choices, and training and inference pipelines. We also describe our broad suite of evaluation metrics.

3.1 Task and data

Task description We generate the main body section (*Findings*) of the report accompanying a chest X-ray. The *Findings* section contains the radiologist's description of the normal and abnormal findings on the image. The image is typically accompanied by an *Indication* section providing the reason for the study, which may include some clinical history or a specific request from the referring clinician.

The most common chest X-ray view is a frontal image, either acquired from posterior to anterior (PA) or anterior to posterior (AP). Other imaging views of the chest are also routinely performed to aid in the diagnostic task, such as a lateral image. It is worth noting that, in isolation, the lateral view is not able to visualise the relevant anatomy with sufficient clarity to generate a comprehensive diagnostic report. Therefore, the frontal view remains the default view for the traditional chest X-ray clinical interpretation task(s) and our work similarly relies on the frontal view in line with most prior studies.[1]

In addition, radiology reports typically contain a final *Impression* section summarising the actionable insights from the study, including suspected clinical diagnoses and recommendations for follow-up investigations. Such information cannot be fully gathered from the image alone—and often not even from the remainder of the report in isolation—relying heavily on the radiologist's domain expertise, external data (e.g. patient history), and context of the study.

Hence, we focus on the task of generating specifically the *Findings* section of the report, given the *Indication* section (if available), and a single frontal chest X-ray image.

Dataset description We train and evaluate on the MIMIC-CXR dataset [Johnson et al., 2019a,b], hosted on PhysioNet [Goldberger et al., 2000]. This dataset from the Beth Israel Deaconess Medical Center in Boston, comprises a total of 377,110 DICOM images across 227,835 studies. Each imaging study is accompanied by a report. We process the DICOM images to remove all non-AP/PA scans. For each report, we extract the *Findings* and *Indication* sections, using the official MIMIC-CXR codebase.[2] We discard all studies for which *Findings* could not be extracted, but we allow for missing *Indication* sections. We use the standard MIMIC-CXR test split in our experiments, however, we additionally exclude benchmark datasets MS-CXR [Boecking et al., 2022] and MS-CXR-T [Bannur et al., 2023b] from our training set. Table 1 describes the subject, study and DICOM counts in each split. To facilitate comparison, we provide the list of DICOM identifiers used in our test split as an ancillary file.

Table 1: Number of subjects, studies and DICOMs in each of our splits of the MIMIC-CXR dataset.

	Train	Validation	Test
Subjects	55,218	2,709	285
Studies	131,613	6,471	2,210
DICOMs	146,909	7,250	2,461

We derive our images from the original DICOM files, instead of the commonly used JPEG files from MIMIC-CXR-JPG [Johnson et al., 2019c], the latter of which may contain compression artefacts and further loss of detail from grayscale quantisation. Following the pre-processing pipeline from CLIP [Radford et al., 2021], images are resized to match the shortest edge to the input size of the image encoder, then the longest edge is centre-cropped to the same size. Intensities are normalised according to the image encoder's training data.

[1] A notable exception is Tu et al. [2023], where the lateral view is used.

[2] https://github.com/MIT-LCP/mimic-cxr/blob/master/txt/section_parser.py

Table 2: Example of the output of GPT-3.5 paraphrasing on the *Indication* and *Findings* sections. Paraphrasing preserves the clinical content while introducing small variations in phrasing.

Original	INDICATION: _F with presyncope. r/o infection // ?pneumonia FINDINGS: AP and lateral chest radiograph demonstrates hyperinflated lungs. Cardiomediastinal and hilar contours are within normal limits. There is no pleural effusion or pneumothorax. No evidence of pulmonary edema. Lungs are without a focal opacity worrisome for pneumonia. There is no air under the right hemidiaphragm.
Paraphrased	INDICATION: Female patient with pre-syncope. Suspected infection, possibly pneumonia. FINDINGS: Chest radiographs show lungs with hyperinflation, but normal cardiomediastinal and hilar contours. No pleural effusion, pneumothorax, or pulmonary edema is observed. No focal opacity is present in the lungs that could indicate pneumonia. Additionally, there is no air under the right hemidiaphragm.

Data augmentation using GPT We use GPT-3.5 to paraphrase both the *Findings* and *Indication* sections of the training set as a data augmentation technique, resulting in an additional 131,558 reports.[3] We instruct GPT-3.5 to rewrite the findings and indication while preserving the information and radiology style. Table 2 shows an example. Note that we leave the validation and test sets unchanged.

3.2 Model architecture

Our model consists of an image encoder, a learnable adapter on top of the image features, and an LLM, following the LLaVA-1.5 architecture [Liu et al., 2023b,a]. We use the radiology-specialised image encoder, RAD-DINO [Pérez-García et al., 2024]. RAD-DINO is a ViT-B model [Dosovitskiy et al., 2020] with 87M parameters trained on 838k chest X-ray images. The input resolution is 518×518 and the encoder patch size is 14. We employ Vicuna-7B [Chiang et al., 2023] as the LLM. We note from LLaVA-Med [Li et al., 2023] that this initialisation from an LLM pretrained only on language should lead to better performance than from one already trained on multimodal data. Our adapter is a multi-layer perceptron (MLP) with GELU activations [Hendrycks and Gimpel, 2016] and hidden size 1024 for all layers. The adapter weights are randomly initialised following the defaults from PyTorch v2.0.1.

We prompt the model with a system message and a human instruction interleaved with the relevant image. The model is trained to output the correct answer, i.e. the full *Findings* section of the report. We first embed the image into a sequence of image patch tokens using the image encoder, taking the embeddings from the last layer of the model and excluding the `CLS` token. The adapter MLP is applied to each embedding to transform these image tokens to the input space of the LLM. After tokenising and embedding the prompt and answer, we insert the image patch tokens at the specified location in the prompt (typically between the system message and human instruction). The instructions we use with MAIRA-1 are "{image placeholder} Provide a description of the findings in the radiology image given the following indication: {indication}" if we know the *Indication* for the image, otherwise "{image placeholder} Provide a description of the findings in the radiology image.".

3.3 Training and inference

We train with a standard auto-regressive language modelling loss (cross-entropy) [Graves, 2013]. We use similar hyperparameters to LLaVA-1.5 for training, i.e. tuning the LLM jointly with the randomly initialized adapter [Liu et al., 2023b]. Unlike LLaVA-1.5 we do not have a precursor training step to pretrain the adapter (see Appx. §A for an experimental comparison). We train for 3 epochs without any parameter-efficient fine-tuning techniques. We use a cosine learning rate scheduler with a warm-up of 0.03 and learning rate 2×10^{-5}. The global batch size is 128. Based on the behaviour of validation metrics observed throughout training in our experiments, we take the final checkpoint for all runs. For inference, we decode in 32-bit precision up to 150 tokens.

3.4 Evaluation metrics

We evaluate the generated reports using both lexical and radiology-specific metrics, as described below. Radiology-specific metrics typically focus on how particular findings are described in the text, for example whether 'consolidation' is noted as present or absent. For this reason, they are more sensitive to clinically-relevant aspects of the generated report, rather than potentially-superficial variation in phrasing.

[3]We used a private, compliant deployment of OpenAI's `gpt-3.5-turbo` version 0301 on Microsoft Azure (https://learn.microsoft.com/en-us/azure/ai-services/openai/overview), such that MIMIC data was not sent to public-facing servers.

Lexical metrics We employ a collection of traditional NLP metrics designed to quantify word overlap between generated and reference texts. In particular, ROUGE-L [Lin, 2004] quantifies the length of the longest common word subsequence relative to the lengths of predicted and reference reports. BLEU-4 [Papineni et al., 2002] is based on n-gram precision (geometric mean for n up to 4), with a brevity penalty to discourage too short predictions. Lastly, we report METEOR [Banerjee and Lavie, 2005], which first aligns individual words (unigrams) in the prediction and reference, while attempting to preserve their ordering; then computes a weighted harmonic mean of unigram precision and recall, with a penalty for fragmentation of consecutive word subsequences.[4]

CheXpert F1 This set of metrics uses the CheXbert automatic labeller [Smit et al., 2020] to extract 'present'/'absent'/'uncertain' labels for each of the 14 CheXpert pathological observations [Irvin et al., 2019] from a generated report and the corresponding reference. As done originally by Irvin et al. [2019], we compute two versions of this metric, mapping the 'uncertain' label to negative or positive, to enable comparison with papers that use either. The binary F1 score for each CheXpert category is then computed conventionally as the unweighted harmonic mean of precision and recall. We report the macro- and micro-averaged F1 scores over 5 major observations[5] and all 14 observations, referring to them respectively as '[Macro/Micro]-F1-[5/14]'.

CheXbert vector similarity This metric feeds the generated and reference reports through the CheXbert model [Smit et al., 2020], then calculates the cosine similarity between their embeddings [Yu et al., 2023]. We compute this metric as well as RadGraph F1 and RadCliQ using the code released by Yu et al. [2023].[6]

RadGraph-based metrics The RadGraph model [Jain et al., 2021] parses radiology reports into graphs containing clinical entities (references to anatomy and observations) and relations between them. Introduced in Yu et al. [2023], the RadGraph F1 metric computes the overlap in entities and relations separately, then reports their average. Entities are considered to match if the text spans and assigned types are the same, and relations are matched if their endpoint entities and relation type are the same.

We also compute a variant F1 metric used in prior work [Delbrouck et al., 2022], to enable direct comparison. Specifically, we report their RG_{ER} score, which matches entities based on their text, type, and whether or not they have at least one relation. For this, we use the `radgraph` package.[7]

RadCliQ Also proposed by Yu et al. [2022], RadCliQ (Radiology Report Clinical Quality) is a composite metric that integrates RadGraph F1 and BLEU score in a linear regression model to predict the total number of errors that radiologists would identify in a report. In their study, this metric had the closest alignment with radiologists' judgement of report quality. For consistency with prior work, we use version 0 of RadCliQ.

4 Experiments

In this section, we present an overview of the experimental design, results and ablations. The main experimental design is seen in Table 3 where we detail design choices related to data, pretraining, image encoders and adapters. To account for slight differences in the definitions of the test split in other studies (see §4.3), for all our experiments we report medians and 95% confidence intervals, estimated over 500 bootstrap samples of the MIMIC-CXR test set. For all metrics, higher values are better, with the exception of RadCliQ (indicated by '↓' in the tables).

4.1 Adapting existing large multimodal models

MAIRA-1 training starts the multimodal alignment from scratch, fully relying on the findings generation task to align the image and text embeddings into a joint representation space. To measure the effect of fine-tuning after aligning on other data, we compare to baselines fine-tuning three existing large multimodal models: LLaVA-1.0 [Liu et al., 2023b] and LLaVA-1.5 [Liu et al., 2023a] in the general domain, and LLaVA-Med [Li et al., 2023] in the biomedical domain.

Table 4 (a) compares the performance of these settings. We note that while fine-tuning LLaVA-1.5 outperforms training a comparable model from scratch (Table 4, 'CLIP+MLP-2'), this advantage is lost when we replace CLIP with a domain-specific image encoder (Table 4, 'RAD-DINO+MLP-2'). Out of the different large multimodal models

[4] We run METEOR with the default parameters as originally presented by Banerjee and Lavie [2005].

[5] Following Miura et al. [2021], Tu et al. [2023], and others, based on the CheXpert 'competition tasks' selected by Irvin et al. [2019], the subset of 5 major categories considered is: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion.

[6] https://github.com/rajpurkarlab/CXR-Report-Metric/tree/v1.1.0

[7] https://pypi.org/project/radgraph/

Table 3: Summary of the experimental settings analysed in this study. 'Continual training' refers to loading the model architecture and parameters as published, then training further on the same dataset as our model, under the same conditions (hyperparameters, prompts, etc.). Note that 'CLIP+MLP-2' is equivalent to LLaVA-1.5 before any LLaVA-1.5 training. 'Findings+' means we train on findings generation with GPT augmentation.

Name	Description	Image encoder	Adapter (init.)	LLM (init.)	Training
LLaVA-1.0-init	Continual training	CLIP-ViT-L-224px	Linear (LLaVA-1.0)	LLaMA-0 (LLaVA-1.0)	Findings
LLaVA-Med-init	Continual training	CLIP-ViT-L-224px	Linear (LLaVA-Med)	LLaMA-0 (LLaVA-Med)	Findings
LLaVA-1.5-init	Continual training	CLIP-ViT-L-336px	MLP-2 (LLaVA-1.5)	Vicuna-7B (LLaVA-1.5)	Findings
CLIP+MLP-2	General domain encoder	CLIP-ViT-L-336px	MLP-2 (Random)	Vicuna-7B	Findings
RAD-DINO +MLP-2	+Use RAD-DINO	RAD-DINO-518px	MLP-2 (Random)	Vicuna-7B	Findings
RAD-DINO +MLP-4	+Increase adapter size	RAD-DINO-518px	MLP-4 (Random)	Vicuna-7B	Findings
MAIRA-1	+Use GPT-augmented data	RAD-DINO-518px	MLP-4 (Random)	Vicuna-7B	Findings+

compared, we find that LLaVA-1.5 performs the best when fine-tuned for the findings generation task, although it lags behind MAIRA-1.

Without such fine-tuning for findings generation, we observed that both LLaVA-1.0/1.5 and LLaVA-Med were incapable of producing meaningful radiology reports, using either the prompt from Li et al. [2023] or our own (see qualitative examples in Fig. 1).

4.2 Which components are most beneficial?

MAIRA-1 differs from LLaVA-based multimodal models in its domain-specific image encoder, a deeper adapter module, and the use of a GPT-augmented dataset during training. Table 3 describes several experiments showing the additive effect of these optimizations included in MAIRA-1, with results presented in Table 4 (b). We start from an architecture mirroring LLaVA-1.5, using a pre-trained CLIP image encoder, a randomly initialized two-layer adapter, and Vicuna-7B as the LLM. We train the adapter and LLM jointly on in-domain radiology data to obtain an initial baseline ('CLIP+MLP-2'). Note that the performance of this model is slightly worse than that of a model fine-tuned from LLaVA-1.5, due to the random initialisation of the adapter.

By switching the image encoder from CLIP to a domain-specific model (RAD-DINO) we observe an improvement across all metrics, more than compensating for the need to re-initialise the adapter. The use of the domain-specific encoder also increases the number of image tokens from 576 to 1369. By correspondingly increasing the size of the adapter from two to four layers ('MLP-4'), we show further improvements across both clinical and lexical metrics.

As a final step, we augment the dataset with GPT-paraphrased samples (§3) to produce MAIRA-1. The effect of the GPT-augmentation is to increase clinical metrics while slightly harming lexical metrics. This could be explained by a mild distribution shift in the GPT-paraphrased reports respective to the original MIMIC reports. In Appx. §B we demonstrate that gains from the use of GPT-paraphrased samples are not simply due to training longer.

4.3 How does MAIRA-1 compare to existing approaches?

Strict comparison with prior work is challenging due to variation in test set inclusion criteria and pre-processing steps, despite the existence of a 'canonical' test split for MIMIC-CXR. For example, Tu et al. [2023] starts from the official split, but includes lateral images paired with the study's report as independent samples, resulting in a reported test size of 4,834 images. Yu et al. [2023] and Jeong et al. [2023] take only a single image for each study, resulting in 1,597 samples in the test set[8], and Tanida et al. [2023] follows the split provided by Chest ImaGenome [Wu et al., 2022]. Recall that our test set size is 2,461 image-report samples. We consider a full reproducibility study in the style of Johnson et al. [2017] out of scope for this work, however to enable future comparison we share the image identifiers used in our test set in an ancillary file. Changes in the distribution of the test set can significantly impact reported numbers, as demonstrated in §4.4. We attempt to account for some of this variability by reporting 95% confidence intervals from bootstrap replicates on the test set when we compare with prior work, but numbers must be interpreted with caution.

Given the above caveats, in Table 5 we compare MAIRA-1 to prior work. For all lexical metrics, MAIRA-1 seemingly outperforms or matches prior state of the art (SOTA). The substantial increase in METEOR relative to Tanida et al. [2023] is notable. On clinical metrics, there is no single superior approach—Tu et al. [2023] report slightly superior

[8]We observed 2,210 studies after taking one image per study.

Table 4: Effect of model design choices on findings generation performance. We report median and 95% confidence intervals based on 500 bootstrap samples from the MIMIC-CXR test set. **(a)** Comparison of our approach to 3 baselines based on continually training LLaVA-style models **(b)** Additive gains from using RAD-DINO, increasing the adapter size, and adding GPT-augmented data. **Bold** indicates best performance across all experiments. '↓' indicates that lower is better (RadCliQ only). CheXpert F1 metrics are computed based on CheXbert labeller ouputs.

Metric	**(a)** Continually trained baselines			**(b)** Additive optimizations			
	LLaVA-1.0-init	LLaVA-Med-init	LLaVA-1.5-init	CLIP +MLP-2	RAD-DINO +MLP-2	RAD-DINO +MLP-4	MAIRA-1
Lexical:							
ROUGE-L	27.9 [27.5, 28.5]	27.6 [27.1, 28.1]	29.2 [28.7, 29.7]	28.2 [27.8, 28.8]	**29.8 [29.2, 30.3]**	**30.1 [29.6, 30.6]**	28.9 [28.4, 29.4]
BLEU-1	35.5 [34.8, 36.1]	35.4 [34.8, 36.0]	35.6 [35.0, 36.3]	32.8 [32.0, 33.5]	35.4 [34.7, 36.2]	37.7 [37.1, 38.4]	**39.2 [38.7, 39.8]**
BLEU-4	**15.0 [14.6, 15.5]**	**14.9 [14.5, 15.3]**	13.9 [13.4, 14.3]	12.7 [12.2, 13.2]	14.1 [13.6, 14.6]	**14.9 [14.4, 15.4]**	14.2 [13.7, 14.7]
METEOR	**35.5 [35.0, 35.9]**	**35.3 [34.8, 35.8]**	31.9 [31.4, 32.5]	30.3 [29.8, 30.9]	32.2 [31.6, 32.7]	33.4 [32.8, 34.0]	33.3 [32.8, 33.8]
Clinical:							
RadGraph-F1	19.9 [19.3, 20.5]	19.1 [18.6, 19.7]	21.5 [20.9, 22.2]	20.3 [19.7, 20.9]	23.0 [22.4, 23.6]	**23.8 [23.2, 24.5]**	**24.3 [23.7, 24.8]**
RG_{ER}	24.4 [23.8, 24.9]	23.8 [23.3, 24.4]	26.4 [25.9, 27.1]	25.0 [24.4, 25.6]	27.8 [27.1, 28.4]	28.8 [28.3, 29.4]	**29.6 [29.0, 30.2]**
CheXbert vector	37.7 [36.8, 38.7]	36.9 [36.0, 37.9]	41.3 [40.4, 42.2]	39.6 [38.7, 40.5]	43.0 [42.1, 43.9]	**43.8 [43.0, 44.6]**	**44.0 [43.1, 44.9]**
RadCliQ (↓)	3.27 [3.23, 3.30]	3.31 [3.28, 3.35]	3.22 [3.18, 3.25]	3.29 [3.25, 3.32]	**3.14 [3.10, 3.17]**	**3.10 [3.06, 3.13]**	**3.10 [3.07, 3.14]**
CheXpert F1, uncertain as negative:							
Macro-F1-14	25.5 [24.2, 26.8]	26.9 [25.5, 28.5]	29.6 [28.3, 31.0]	29.4 [27.7, 30.8]	33.4 [31.8, 34.9]	36.6 [35.0, 38.3]	**38.6 [37.1, 40.1]**
Micro-F1-14	43.6 [42.3, 44.7]	42.7 [41.5, 44.0]	49.0 [47.8, 50.3]	46.4 [45.0, 47.6]	52.2 [51.1, 53.4]	54.6 [53.4, 55.8]	**55.7 [54.7, 56.8]**
Macro-F1-5	36.5 [34.7, 38.2]	36.3 [34.4, 38.1]	41.7 [39.8, 43.8]	40.7 [38.8, 42.6]	43.2 [41.7, 45.0]	**46.0 [44.3, 48.1]**	**47.7 [45.6, 49.5]**
Micro-F1-5	45.2 [43.4, 46.8]	43.9 [42.2, 45.6]	50.5 [48.9, 52.3]	48.1 [46.6, 49.8]	53.6 [52.2, 55.2]	**55.2 [53.8, 56.8]**	**56.0 [54.5, 57.5]**
CheXpert F1, uncertain as positive:							
Macro-F1-14+	29.6 [28.5, 30.6]	30.6 [29.3, 32.1]	34.2 [33.0, 35.5]	33.0 [31.6, 34.4]	37.9 [36.6, 39.3]	40.4 [38.9, 41.9]	**42.3 [40.9, 43.6]**
Micro-F1-14+	44.5 [43.4, 45.5]	43.7 [42.5, 44.8]	49.6 [48.4, 50.7]	46.8 [45.7, 48.0]	52.6 [51.5, 53.7]	**54.5 [53.5, 55.6]**	**55.3 [54.3, 56.2]**
Macro-F1-5+	41.5 [39.9, 43.0]	41.4 [39.6, 43.1]	46.8 [45.3, 48.6]	45.6 [43.9, 47.2]	49.4 [47.8, 51.1]	**51.1 [49.4, 52.8]**	**51.7 [49.9, 53.1]**
Micro-F1-5+	48.5 [47.0, 49.8]	47.1 [45.4, 48.7]	54.0 [52.4, 55.6]	51.5 [50.1, 53.0]	56.7 [55.4, 58.3]	**58.1 [56.6, 59.5]**	**58.8 [57.4, 60.0]**

scores on RadGraph-F1 while Delbrouck et al. [2022] substantially outperforms on RG_{ER}—this is perhaps expected given their model was optimised for RG_{ER}. Across the CheXbert-derived metrics, MAIRA-1 is comparable or superior across the fourteen-classes, but slightly underperforms on the five-class subset relative to Tu et al. [2023]. We present class-stratified results for MAIRA-1 in §4.4 (Table 6). Promisingly across all clinical metrics, MAIRA-1 sets a new standard for the radiologist-aligned RadCliQ score.

4.4 Stratified results

Performance depends on finding class Table 6 shows a breakdown of MAIRA-1 performance based on the finding class. We use the standard 14 hierarchical classes initially proposed by Irvin et al. [2019] and used in the CheXbert labeller [Smit et al., 2020].

Looking at F_1-score, we note that the Macro-F1-14 values we report elsewhere mask variance across classes. We see high-performing classes (Support Devices: 84.5, Pleural Effusion: 68.9, Cardiomegaly: 64.0) and indeed poorly performing classes (Enlarged Cardiomediastinum: 11.9, Pleural Other: 14.7, Pneumonia: 18.3). We note that these latter categories are rarer and more nebulously defined,[9] and may be subject to more noise in the output of the CheXbert labeller itself. For example, Cardiomegaly (37% prevalence) should imply an enlarged cardiomediastinum, and yet the latter category occurs in only 8% of cases.

Specificity and negative predictive value (NPV) are not measured in the Macro-F1-14 aggregate metric, where we see consistently high values. Compared to recall and precision (positive predictive value), we infer the model may under- or miss-call positive findings, but more reliably reports on the absence of findings. This is useful for findings such as pneumothorax (NPV 98.9) in an intensive care setting, where a chest X-ray may serve to confirm the *absence* of a pneumothorax after the insertion of a drain or tube.

Results differ between normal and abnormal studies When there are no findings in a study, radiology reports are often formulaic, containing templated phrases such as "No [evidence of] acute cardiopulmonary process". This

[9] As acknowledged by Irvin et al. [2019], pneumonia is a clinical diagnosis which should strictly not be assessed from a chest X-ray alone.

Table 5: Findings generation performance on the MIMIC-CXR test set, compared to closest state-of-the-art for each metric. We report median with 95% confidence intervals from 500 bootstrap samples of the test-set for MAIRA-1. Prior work uses differing test sets, limiting comparison. Our model is composed of a 86.6M parameter image encoder, 53M MLP adapter and 7B LLM. Our test size is 2461 samples. For the CheXpert F1 metrics, '+' means the uncertain class is mapped to positive, otherwise it is mapped to negative. *The CheXbert vector score is an evaluation of the Miura et al. [2021] model performed by Yu et al. [2023]. †The RadCliQ number is an evaluation of Miura et al. [2021] model reported by Jeong et al. [2023], Yu et al. [2023], using RadCliQ-v0.

Category	Metric	MAIRA-1	SOTA [ref.]	Param. count Vision / LLM	Test set size
Lexical	ROUGE-L	**28.9** [28.4, 29.4]	27.49 [Tu et al., 2023]	22B / 62B	4,834 images
	BLEU-1	**39.2** [38.7, 39.8]	32.31 [Tu et al., 2023]	22B / 62B	4,834 images
	BLEU-4	**14.2** [13.7, 14.7]	13.30 [Miura et al., 2021]	8M / ?	2,347 reports
	METEOR	**33.3** [32.8, 33.8]	16.8 [Tanida et al., 2023]	26M / 355M	32,711 images
Clinical	RadGraph-F1	24.3 [23.7, 24.8]	**26.71** [Tu et al., 2023]	22B / 62B	4,834 images
	RG_{ER}	29.6 [29.0, 30.2]	**34.7** [Delbrouck et al., 2022]	8M / 18M	2,347 reports
	CheXbert vector	44.0 [43.1, 44.9]	**45.2** [Miura et al., 2021]*	8M / ?	1,597 images
	RadCliQ (↓)	**3.10** [3.07, 3.14]	3.277 [Miura et al., 2021]†	8M / ?	1,597 images
	Macro-F1-14	**38.6** [37.1, 40.1]	**39.83** [Tu et al., 2023]	22B / 62B	4,834 images
	Micro-F1-14	**55.7** [54.7, 56.8]	53.56 [Tu et al., 2023]	22B / 62B	4,834 images
	Macro-F1-5	47.7 [45.6, 49.5]	**51.60** [Tu et al., 2023]	22B / 62B	4,834 images
	Micro-F1-5	56.0 [54.5, 57.5]	**57.88** [Tu et al., 2023]	22B / 62B	4,834 images
	Macro-F1-14+	42.3 [40.9, 43.6]	–		
	Micro-F1-14+	55.3 [54.3, 56.2]	–		
	Macro-F1-5+	51.7 [49.9, 53.1]	–		
	Micro-F1-5+	**58.8** [57.4, 60.0]	54.7 [Tanida et al., 2023]	26M / 355M	32,711 images

constitutes a simpler language generation task for the model, as reflected in the higher lexical metrics in Table 7. In addition, we observe higher clinical metrics in the no-finding subset, which we speculate could be related to the distribution shift between MIMIC-CXR dataset splits. As noted by Johnson et al. [2019c], studies with no findings are over-represented in the released training and validation splits, compared to the test set.

The model benefits from the indication section The study indication is expected to substantially affect the contents of a radiology report. For example, as in Table 2, it may prompt the radiologist to report on specific types of abnormality that might not be routinely included, and is particularly relevant for deciding what to report as absent from the scan. MAIRA-1 uses the indication section of the report to help generate the findings section, whenever it is available with the study (66.3% in training and 57.5% in the test set). In Table 7 we show how performance varies in subsets of the test set with and without the indication section. Indeed, by leveraging the indication, MAIRA-1 is able to to generate radiology reports that are drastically more similar and more accurate with respect to the true report than in the subset without indication.

Table 6: Breakdown of metrics per CheXpert finding class, as defined by Irvin et al. [2019]. Classes are hierarchical and not mutually exclusive. The positive class is 'present', and 'uncertain' is mapped to negative. 'Lung Lesion' includes masses, nodular densities or opacities, lumps, and tumors. 'Pleural Other' includes pleural or parenchymal thickening or scarring, as well as fibrosis. 'Support Devices' includes lines, tubes, catheters, pacemakers, coils, drains, etc. NPV = negative predictive value. Performance numbers are median and 95% confidence intervals from 500 bootstrap replicates from the MIMIC-CXR test set.

Finding class	% (n, median)	Precision	Recall	NPV	Specificity	F_1-score
No Finding	6% (151)	31.6 [26.3, 37.9]	49.1 [41.7, 56.8]	96.6 [95.8, 97.3]	93.1 [92.1, 94.0]	38.6 [32.6, 44.5]
Lung Opacity	38% (944)	58.0 [54.6, 61.6]	43.7 [40.5, 46.6]	69.6 [67.7, 71.9]	80.3 [78.2, 82.2]	49.8 [47.1, 52.7]
Atelectasis	28% (688)	43.3 [39.5, 47.3]	39.4 [35.6, 43.4]	77.3 [75.4, 79.1]	79.9 [78.0, 81.8]	41.3 [37.8, 44.6]
Edema	18% (436)	47.4 [42.1, 52.6]	41.2 [37.1, 45.9]	87.7 [86.4, 89.0]	90.1 [88.6, 91.4]	44.0 [39.9, 48.6]
Lung Lesion	6% (146)	30.1 [18.4, 41.6]	13.6 [7.9, 18.6]	94.7 [93.9, 95.5]	98.0 [97.4, 98.5]	18.8 [11.4, 24.9]
Consolidation	5% (115)	25.9 [16.3, 36.2]	16.4 [10.2, 23.8]	96.0 [95.1, 96.8]	97.7 [97.0, 98.3]	20.0 [12.9, 28.1]
Pneumonia	5% (113)	22.1 [13.7, 32.4]	15.5 [9.6, 22.8]	96.0 [95.2, 96.9]	97.3 [96.7, 98.0]	18.3 [11.7, 25.5]
Cardiomegaly	37% (903)	61.7 [58.7, 64.5]	66.3 [63.2, 69.5]	79.6 [77.6, 81.8]	76.2 [74.1, 78.1]	64.0 [61.2, 66.6]
Enlarged Cardiomediastinum	8% (196)	13.2 [8.1, 18.6]	10.6 [6.8, 15.2]	92.4 [91.4, 93.5]	93.9 [92.9, 94.9]	11.9 [7.4, 16.4]
Pleural Effusion	34% (833)	69.0 [65.9, 72.2]	68.6 [65.3, 72.0]	83.9 [82.1, 86.0]	84.3 [82.3, 85.9]	68.9 [66.3, 71.3]
Pleural Other	3% (77)	23.7 [10.1, 38.4]	10.8 [4.3, 18.2]	97.2 [96.6, 97.8]	98.9 [98.5, 99.3]	14.7 [6.2, 23.5]
Pneumothorax	2% (55)	34.2 [24.4, 44.4]	50.8 [37.4, 65.5]	98.9 [98.4, 99.3]	97.8 [97.2, 98.3]	40.8 [29.7, 51.2]
Fracture	5% (130)	37.0 [25.9, 50.0]	18.8 [12.3, 25.8]	95.6 [94.8, 96.4]	98.2 [97.7, 98.7]	24.9 [17.3, 33.1]
Support Devices	41% (1001)	84.6 [82.4, 86.7]	84.4 [82.0, 86.4]	89.3 [87.7, 90.8]	89.4 [87.9, 91.1]	84.5 [82.7, 86.0]

Table 7: Breakdown of MAIRA-1 metrics by (i) whether the study has the CheXpert label 'No finding' and (ii) whether the report contains an indication for the study. (i) For 'normal' cases (no finding), the original reports tend to be more formulaic—which could explain the higher lexical metrics in this group—as well as defining a somewhat simpler clinical task for the model. (ii) The results are strikingly superior for cases that contain an indication, as the model is able to leverage the strong cues for what findings (positive or negative) should be reported. All performance numbers are median and 95% confidence intervals from 500 bootstrap replicates from the MIMIC-CXR test set.

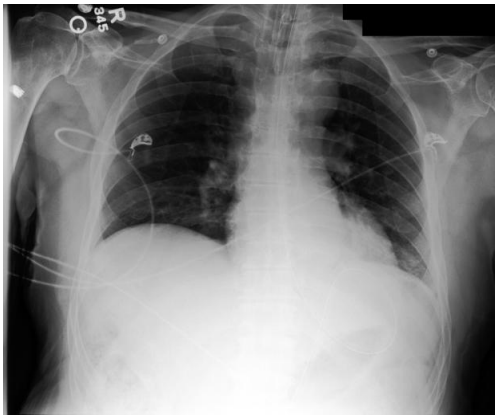
Category	Metric	Has finding	No finding	Has indication	No indication
	% (n)	78.3% (1928)	21.7% (533)	57.5% (1414)	42.5% (1047)
Lexical	ROUGE-L	27.6 [27.1, 28.1]	**33.4 [31.9, 34.9]**	**32.7 [32.0, 33.4]**	23.6 [23.1, 24.1]
	BLEU-4	13.2 [12.7, 13.7]	**18.7 [17.4, 20.3]**	**17.6 [16.9, 18.2]**	9.0 [8.6, 9.6]
	METEOR	31.9 [31.4, 32.5]	**38.5 [37.1, 40.0]**	**36.8 [36.1, 37.6]**	28.6 [28.0, 29.3]
Clinical	RadGraph-F1	23.0 [22.4, 23.5]	**28.5 [27.0, 29.9]**	**27.8 [27.0, 28.5]**	19.2 [18.6, 20.0]
	RG_{ER}	28.5 [27.9, 29.0]	**33.7 [32.0, 35.3]**	**33.5 [32.8, 34.3]**	24.3 [23.6, 25.0]
	ChexBert vector	42.3 [41.3, 43.2]	**50.3 [48.1, 52.2]**	**47.3 [46.2, 48.3]**	39.5 [38.2, 40.8]
	RadCliQ (↓)	3.19 [3.16, 3.22]	**2.79 [2.71, 2.88]**	**2.88 [2.84, 2.92]**	3.41 [3.37, 3.45]
	Macro-F1-14	38.1 [36.2, 39.9]	–	39.1 [37.1, 41.5]	36.9 [34.4, 39.2]
	Micro-F1-14	57.0 [55.8, 57.9]	–	56.3 [54.7, 57.8]	55.0 [53.3, 56.6]
	Macro-F1-5	48.8 [46.7, 50.8]	–	46.8 [44.5, 50.0]	47.3 [44.4, 50.4]
	Micro-F1-5	57.5 [56.0, 59.0]	–	56.6 [54.6, 58.6]	55.5 [53.3, 57.8]

5 Examples

Fig. 1 reproduces the example[10] shown in Tu et al. [2023], comparing the output of MAIRA-1 with the best Med-PaLM M variant (84B) and the closest by model size (12B). For illustration, we also include the output of models not trained on the findings generation task (LLaVA-1.5 and LLaVA-Med).

To conduct further error analysis, we sorted all examples in the test set by their RadCliQ score and manually reviewed the top and bottom cases with two consultant radiologists. The highest-scoring examples (lowest RadCliQ) were largely cases without findings, where the report follows a standard format; this is seen already in Table 7 when we stratify by

[10] In this section we show the JPEG versions of images as released in MIMIC-CXR-JPG [Johnson et al., 2019c]. However, as described in §3, MAIRA-1 ingests images derived from the original DICOMs.



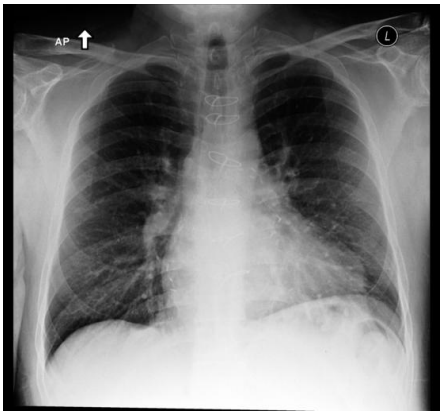
Original report

INDICATION: _-year-old male with intubation. No prior examinations for comparison.

FINDINGS: Endotracheal tube terminates 6 cm from the carina. Nasogastric tube loops within the stomach, with side port within the gastric body. There are no significant pleural effusions or pneumothorax. Mild retrocardiac atelectasis is present. Heart size is normal. The aorta is tortuous.

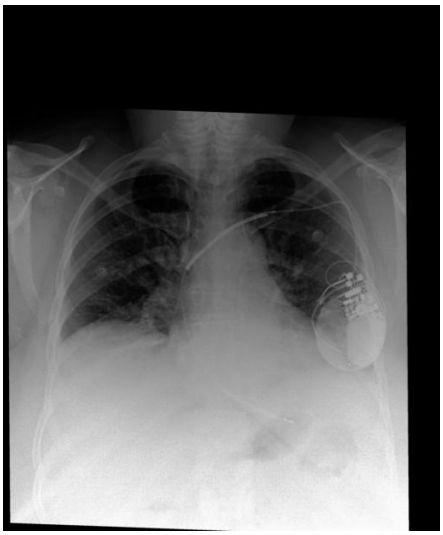
Model	Generated findings
MAIRA-1	Endotracheal tube terminates 5 cm above the carina. Nasogastric tube loops in the stomach, with side port near the GE junction. Heart size is normal. Aorta is tortuous and unfolded. There are no pleural effusions or pneumothorax. Lungs are clear. [*No comment about atelectasis.*]
Med-PaLM M-12B	An endotracheal tube terminates 5 cm above the carina. There is no pneumothorax or large pleural effusion. Opacities in the medial right lower lobe and retrocardiac region may represent atelectasis aspiration or pneumonia. A calcified lymph node is noted in the aortopulmonary window. Cardiac size is normal. [*No comment about NG/OG tube.*]
Med-PaLM M-84B	Endotracheal tube terminates 5 cm above the carina. Nasogastric tube courses below the diaphragm and inferiorly beyond the film. There is retrocardiac opacity which may represent atelectasis or aspiration. The heart size is normal. The aorta is tortuous and unfolded. No pleural effusions or pneumothorax.
LLaVA-Med	The image is a chest X-ray taken on day 1 post-admission. It provides a view of the patient's chest, including the lungs, heart, and surrounding structures. Chest X-rays are commonly used to help diagnose and monitor various conditions affecting the respiratory and cardiovascular systems, such as infections, fluid accumulation, or structural abnormalities.
LLaVA-1.5	The image displays a close-up view of a human body, focusing on the chest area. The chest is filled with various medical instruments, including a heart monitor and a defibrillator. The heart monitor is placed on the left side of the chest, while the defibrillator is positioned on the right side. In addition to the medical instruments, there are several wires and cords scattered throughout the scene. Some of these wires are attached to the medical devices, while others are likely part of the electrical connections for the instruments. The presence of these wires and cords suggests that the medical equipment is in use, possibly for monitoring or treating a patient.

Figure 1: Model outputs on the sample from Tu et al. [2023]. MAIRA-1 and Med-PaLM M-84B produce similar reports with different errors; MAIRA-1 misses the retrocardiac atelectasis and incorrectly locates the side port of the nasogastric tube, whereas Med-PaLM M-84B incorrectly describes the NG tube as ending beyond the film. For comparison purposes, we include Med-PaLM M-12B, which hallucinates a calcified lymph node and fails to mention the NG tube, and the outputs from LLaVA-Med and LLaVA-1.5, which do not produce realistic radiology reports.



Original report	INDICATION: _-year-old man with cough and fever, question acute process. FINDINGS: The lungs are well expanded and clear. The cardiomediastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.
MAIRA-1	FINDINGS: The lungs are well expanded and clear. The cardiomediastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Median sternotomy wires and surgical clips are noted.

Figure 2: MAIRA-1 successfully generates the *Findings* section for a normal chest X-ray. Highlighted is a small variation in how the sternal wires are described; here the original report is slightly more specific, but MAIRA-1 is not incorrect. RadCliQ: 0.239.



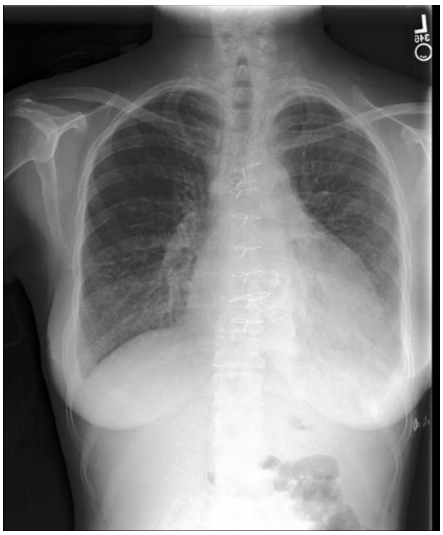
Original report	INDICATION: _F with shortness of breath, congestive heart failure. FINDINGS: Lung volumes are low compared to the previous study. Left-sided AICD device is noted with single lead terminating in unchanged position in the right ventricle. Heart size appears at least mildly enlarged. The mediastinal and hilar contours are unremarkable. There is crowding of the bronchovascular structures without overt pulmonary edema. Streaky opacities in the lung bases likely reflect areas of atelectasis. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.
MAIRA-1	FINDINGS: Lung volumes are low. Left-sided AICD device is noted with single lead terminating in the region of the right ventricle. Heart size is mildly enlarged. Mediastinal and hilar contours are unremarkable. Crowding of bronchovascular structures is present without overt pulmonary edema. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

Figure 3: MAIRA-1 generates the *Findings* section for a more complex case with multiple findings. Here, the original report describes the lung volumes in relation to a prior image which is not available to MAIRA-1; in this case MAIRA-1 simply describes the lung volumes as low (highlighted). RadCliQ: 0.239.

whether the case has a finding. Fig. 2 shows such an example; here MAIRA-1 has generated an almost-identical report to that of the original radiologist, with the exception of providing less detail on the sternal wires.

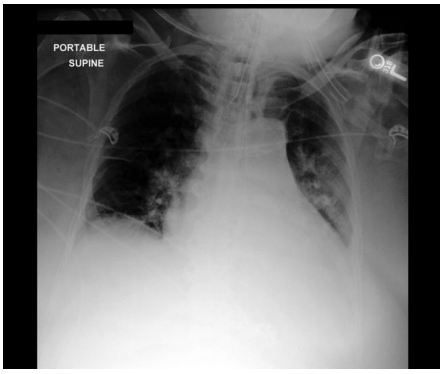
A recurring but expected 'failure' mode we observe in MAIRA-1 is the generation of descriptions of change. This is a well-known problem in the generation of radiology reports [Bannur et al., 2023a, Ramesh et al., 2022]: Standard clinical practice involves reporting using both current and prior studies precisely to enable assessment of change. This version of MAIRA-1 is trained on reports from single images, where reports often contain descriptions of changes compared to a previous study. This type of report is then also sometimes produced at test time, as illustrated in some of the following examples.

Figs. 3 and 4 show success cases from MAIRA-1. In Fig. 3, MAIRA-1 successfully *avoids* hallucinating a comparison to a prior study ('Lung volumes are low.'), while Fig. 4 shows examples of such hallucination ('... pulmonary edema, new compared to the prior study.'), with other details largely correct.



Original report	INDICATION: *[No indication]* FINDINGS: The patient is status post median sternotomy, CABG, and mitral valve replacement. The heart is mildly enlarged. The mediastinal contours are unchanged with calcification of the aortic knob again noted. Mild pulmonary edema appears progressed compared to the prior exam with small bilateral pleural effusions, also minimally increased compared to the prior exam. Left basilar opacification likely reflects atelectasis. There is no pneumothorax. No acute osseous abnormalities are identified.
MAIRA-1	FINDINGS: The patient is status post median sternotomy and mitral valve replacement. Heart size is moderately enlarged but unchanged. The mediastinal contours are stable. There is mild interstitial pulmonary edema, new compared to the prior study. Small bilateral pleural effusions are also new. Retrocardiac opacity likely reflects atelectasis. There is no pneumothorax. No acute osseous abnormalities are visualized.

Figure 4: MAIRA-1 generates the *Findings* section of a complex case. Here, although MAIRA-1 correctly describes the presence and absence of findings, it generates descriptions of change which cannot be justified from a single image. RadCliQ: 1.053.



Original report	INDICATION: *[No indication]* FINDINGS: Compared to the prior exam there is no significant interval change.
MAIRA-1	FINDINGS: The endotracheal tube is positioned 4 cm above the carina, while the nasogastric tube is located below the diaphragm. The cardiac silhouette is enlarged with pulmonary vascular congestion. The left base shows opacification, indicating volume loss in the lower lobe and pleural effusion. The right base shows less prominent atelectatic changes and a possible small effusion.

Figure 5: The original *Findings* section in this report does not describe any of the observed findings, comparing only to the previous image. Here, MAIRA-1 correctly describes the findings in the image. The RadCliQ score (5.327) is high due to low overlap between the reports, unfairly penalising the model's output.

Fig. 5 shows a case with high (poor) RadCliQ score, however this does not reflect a failure of the model and rather a limitation of the evaluation. In this case, the original report contains negligible detail ('Compared to the prior exam there is no significant interval change.'), whereas MAIRA-1 generates a full report which is largely correct. We note that examples of the converse also exist, where MAIRA-1 simply refers to the absence of change without further elaboration (for 491 studies, the *Findings* section is exactly the sentence 'Compared to the prior study there is no significant interval change.'). This underscores a limitation of training with and evaluating on 'noisy' real-world datasets.

We further note that, in Fig. 5, MAIRA-1 has generated a quantitative measurement ('... tube is positioned 4cm above the carina'), which is not grounded in physical measurements of the image. This is also observed in Fig. 1 from both MAIRA-1 and Med-PaLM M. Whereas a model may learn about an average field-of-view of the images seen during training, as well as certain correlations in the training reports, such measurements cannot be produced accurately without knowledge of physical and geometric parameters of the image acquisition.

6 Discussion

We have presented results on MAIRA-1, a radiology-specialised large multimodal model designed for the task of generating the *Findings* section of a chest X-ray report.

Architecturally, MAIRA-1 consists of a frozen domain-specific image encoder based on a ViT-B (RAD-DINO), a four-layer feedforward adapter module, and the LLM Vicuna-7B. We train MAIRA-1 solely on the open-access MIMIC-CXR dataset, leveraging GPT-3.5 for text-based data augmentation. With these components, we demonstrate that performance competitive with existing state-of-the-art is possible, with either fewer parameters or a simpler training objective.

The use of a domain-specific image encoder significantly boosts the performance of MAIRA-1. One aspect of this may be the increased image resolution of RAD-DINO (518px), enabling the detection of potentially small and/or subtle findings such as pneumothorax. A larger set of image tokens enabled us to fruitfully scale up the adapter layer, raising the prospect of further gains from yet more complex processing of image tokens.

By using GPT-3.5 to paraphrase reports in MIMIC-CXR, we observe a boost in clinical metrics with a minor degradation of lexical metrics. We speculate that this paraphrasing serves as a semantics-preserving transformation of the text, encouraging the model to focus on the key aspects of the report without overfitting to its style.

Our stratified analysis reveals significantly higher performance on studies containing an *Indication* section. We hypothesise two mechanisms for this. Firstly, knowledge of the 'question' behind the report may inform the expected *Findings*; for example, '...please evaluate for pleural effusion' (*Indication*) prompts: 'Left mid to lower lung opacified is likely a combination of moderate right-sided pleural effusion...' (*Findings*). Secondly, the *Indication* section can include additional clinical context on the patient, which has been shown to improve interpretation [Yapp et al., 2022].

By stratifying by findings class, we observe that the behaviour of MAIRA-1 varies, yielding the best overall performance on classes such as support devices, pleural effusion, and cardiomegaly. For certain clinically actionable findings such as edema and consolidation, MAIRA-1 exhibits unsatisfactory recall, albeit with consistently high negative predictive value. This highlights that reporting aggregate metrics alone may obscure disparate performance within findings classes or patient subgroups.

To understand whether models such as MAIRA-1 are clinically useful, more fine-grained metrics, categories, and exemplar datasets are important, as well as evaluations in realistic use-contexts [Huang et al., 2023]. For example, the commonly used CheXpert classes include 'Pneumonia' despite it being a clinical diagnosis which should not be assessed from an image alone. Efforts such as RadCliQ are valuable towards the development of radiology-specific evaluation methods, but, as highlighted in our error analysis, the presence of incomplete or otherwise 'imperfect' reports in the dataset mean correctly generated reports may still be unfairly penalised.

Reports often contain information derived from prior studies, clinical notes, accompanying laterals or other relevant imaging examinations of the patient. Models such as MAIRA-1 and much prior work, trained to generate reports from single images, are thus forced to hallucinate this information from their limited context. Existing commonly reported metrics do not directly quantify this failure mode. Future versions of MAIRA-1 could include the current and previous study, thereby reducing the need to hallucinate, as demonstrated in Bannur et al. [2023a].

7 Conclusion

We have presented MAIRA-1 as a proof-of-concept for a radiology-adapted large multimodal model. Despite training on a relatively small dataset with a conventional language-modelling loss, it exhibits competitive performance with existing state-of-the-art on findings generation across a broad suite of metrics, benefiting from a domain-specific image encoder, simple training paradigm and text-based augmentation. We believe the performance and clinical utility of MAIRA-1 can be pushed much further by allowing the model to consider multiple images, e.g. priors and complementary views, and by training on larger, more diverse, and higher-quality datasets.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay,

Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 technical report, 2023. URL https://arxiv.org/abs/2305.10403.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo: An open-source framework for training large autoregressive vision-language models, August 2023. URL http://arxiv.org/abs/2308.01390. arXiv:2308.01390 [cs].

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*, pages 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*, pages 15016–15027, 2023a.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Max Ilse, Daniel Coelho de Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anton Schwaighofer, Maria Teodora Wetscherek, Hannah Richardson, Tristan Naumann, Javier Alvarez Valle, and Ozan Oktay. MS-CXR-T: Learning to exploit temporal structure for biomedical vision-language processing (version 1.0.0), 2023b. URL https://physionet.org/content/ms-cxr-t/1.0.0/.

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/.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. URL https://aclanthology.org/2020.emnlp-main.112/.

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/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

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*, pages 4348–4360. ACL, December 2022. doi:10.18653/v1/2022.findings-emnlp.319.

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.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth,

Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pages 8469–8488, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, page 209–219. PMLR, November 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Alex Graves. Generating sequences with recurrent neural networks, 2013. URL https://arxiv.org/abs/1308.0850.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. URL https://arxiv.org/abs/1606.08415.

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

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, pages 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.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation, August 2023. URL http://arxiv.org/abs/2308.12604. arXiv:2308.12604 [cs].

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376. PMLR, 2017.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019a. doi:10.1038/s41597-019-0322-0.

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

Alistair E. W. 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, November 2019c. URL http://arxiv.org/abs/1901.07042. arXiv:1901.07042 [cs, eess].

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day, 2023. URL http://arxiv.org/abs/2306.00890.

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

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 Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR, 09–10 Aug 2019. URL https://proceedings.mlr.press/v106/liu19a.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a. URL http://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL http://arxiv.org/abs/2304.08485.

Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Xiang Li, Quanzheng Li, and Tianming Liu. Radiology-GPT: A large language model for radiology, June 2023c. URL http://arxiv.org/abs/2306.08666. arXiv:2306.08666 [cs].

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304. ACL, June 2021. doi:10.18653/v1/2021.naacl-main.416.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-Flamingo: a multimodal medical few-shot learner, July 2023. URL http://arxiv.org/abs/2307.15189. arXiv:2307.15189 [cs].

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting, July 2023. URL http://arxiv.org/abs/2201.09405. arXiv:2201.09405 [cs].

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*, pages 311–318. Association for Computational Linguistics, July 2002. doi:10.3115/1073083.1073135.

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 Teodora 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 138 of *PMLR*, pages 8748–8763, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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*, pages 456–473. PMLR, 2022.

Andrew B. Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Florencia Garcia-Vicente, David Melnick, Yun Liu, Krish Eswaran, Daniel Tse, Neeral Beladia, Dilip Krishnan, and Shravya Shetty. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022. doi:10.1148/radiol.212482.

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)*, pages 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*, pages 7433–7442, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tanida_Interactive_and_Explainable_Region-Guided_Radiology_Report_Generation_CVPR_2023_paper.html.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model, March 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI, July 2023. URL http://arxiv.org/abs/2307.14334. arXiv:2307.14334 [cs].

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_TieNet_Text-Image_Embedding_CVPR_2018_paper.html.

Joy T. Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G. Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2022. URL https://openreview.net/forum?id=Hd5634yVi.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders, August 2023. URL http://arxiv.org/abs/2308.01317. arXiv:2308.01317 [cs, eess].

Benjamin Yan, Ruochen Liu, David E. Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P. O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. Style-aware radiology report generation with RadGraph and few-shot prompting, October 2023. URL http://arxiv.org/abs/2310.17811. arXiv:2310.17811 [cs].

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.

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. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9), September 2023. doi:10.1016/j.patter.2023.100802.

Appendix

A Do we need to pre-train the adapter?

Prior work conducted an initial training of just the adapter layer, keeping the LLM frozen [Liu et al., 2023b, Li et al., 2023]. In Table 8, we investigate this choice. We start from a baseline model with CLIP-ViT-L-336px as image encoder and Vicuna-7B as LLM. We obtain a fine-tuned adapter by training on the findings generation task. We then compare the impact of using this adapter against a randomly initialised adapter of the same size in our typical training setup. Although prior work often uses this two-stage training process, we find that initializing from the pretrained adapter causes the final performance to be significantly worse.

Table 8: Compare starting from a pretrained adapter vs a random one. The pretrained adapter is obtained from a separate training of the image model, adapter and LLM on the findings generation task, freezing both the LLM and the image encoder. Performance numbers are median and 95% confidence intervals from 500 bootstrap replicates from the MIMIC-CXR test set.

Category	Metric	Random adapter init	Pretrained adapter init
Lexical	ROUGE-L	28.2 [27.8, 28.8]	27.4 [27.0, 27.9]
	BLEU-1	32.8 [32.0, 33.5]	31.6 [30.9, 32.3]
	BLEU-4	12.7 [12.2, 13.2]	12.1 [11.7, 12.7]
	METEOR	30.3 [29.8, 30.9]	29.2 [28.6, 29.7]
Clinical	RadGraph-F1	20.2 [19.7, 20.8]	18.4 [17.8, 18.9]
	RG_{ER}	25.0 [24.4, 25.6]	23.0 [22.5, 23.7]
	CheXbert vector	39.6 [38.8, 40.4]	35.5 [34.6, 36.5]
	RadCliQ (↓)	3.29 [3.25, 3.32]	3.42 [3.39, 3.46]
	Macro-F1-14	29.4 [27.7, 30.8]	24.0 [22.6, 25.4]
	Micro-F1-14	46.4 [45.0, 47.6]	39.7 [38.4, 41.0]
	Macro-F1-5	40.7 [38.8, 42.6]	32.6 [30.9, 34.3]
	Micro-F1-5	48.1 [46.6, 49.8]	40.1 [38.2, 41.9]

B Are gains from GPT-augmentation due to training longer?

We see from Table 4 that adding GPT augmentation improves the performance of the model on clinical metrics. Hoewever, adding GPT paraphrased reports also increases the number of samples in the dataset and thus the training steps, as we train for the full 3 epochs even after adding the paraphrased example to our dataset. To examine whether the gains we get from using GPT-augmented data are purely due to training for a larger number of steps, we run an ablation. We create a variant dataset with the same set of samples as that from GPT-augmentation, but using the *original* report rather than the GPT-paraphrased variant. In Table 9 we compare this setting with the use of GPT-paraphrased reports.

Table 9: Experiment controlling for the relationship between training steps and model performance. 'Control' indicates an experiment where we replace the GPT augmented reports with the original in the dataset, thus keeping the training steps constant and changing only the data. Performance numbers are median and 95% confidence intervals from 500 bootstrap replicates from the MIMIC-CXR test set.

Category	Metric	MAIRA-1	Control
Lexical	ROUGE-L	28.9 [28.4, 29.4]	28.5 [28.0, 28.9]
	BLEU-4	14.2 [13.7, 14.7]	14.0 [13.6, 14.5]
	METEOR	33.3 [32.8, 33.8]	32.7 [32.2, 33.2]
Clinical	RadGraph-F1	**24.3** [23.7, 24.8]	22.8 [22.1, 23.3]
	RG_{ER}	**29.6** [29.0, 30.2]	27.8 [27.3, 28.4]
	CheXbert vector	**44.0** [43.1, 44.9]	42.8 [42.0, 43.6]
	RadCliQ (↓)	**3.10** [3.07, 3.14]	3.16 [3.13, 3.19]
	Macro-F1-14	38.6 [37.1, 40.1]	36.7 [35.0, 38.2]
	Micro-F1-14	**55.7** [54.7, 56.8]	54.3 [53.3, 55.4]
	Macro-F1-5	**47.7** [45.6, 49.5]	45.2 [43.3, 47.1]
	Micro-F1-5	56.0 [54.5, 57.5]	54.4 [53.0, 55.8]