

Advancing Multimodal Medical Capabilities of Gemini

Google Research and Google DeepMind [†]

Many clinical tasks require an understanding of specialized data, such as medical images and genomics, which is not typically found in general-purpose large multimodal models. Building upon Gemini's multimodal models, we develop several models within the new *Med-Gemini* family that inherit core capabilities of Gemini and are optimized for medical use via fine-tuning with 2D and 3D radiology, histopathology, ophthalmology, dermatology and genomic data. Med-Gemini-2D sets a new standard for AI-based chest X-ray (CXR) report generation based on expert evaluation, exceeding previous best results across two separate datasets by an absolute margin of 1% and 12%, where 57% and 96% of AI reports on normal cases, and 43% and 65% on abnormal cases, are evaluated as "equivalent or better" than the original radiologists' reports. We demonstrate the first ever large multimodal model-based report generation for 3D computed tomography (CT) volumes using Med-Gemini-3D, with 53% of AI reports considered clinically acceptable, although additional research is needed to meet expert radiologist reporting quality. Beyond report generation, Med-Gemini-2D surpasses the previous best performance in CXR visual question answering (VQA) and performs well in CXR classification and radiology VQA, exceeding SoTA or baselines on 17 of 20 tasks. In histopathology, ophthalmology, and dermatology image classification, Med-Gemini-2D surpasses baselines across 18 out of 20 tasks and approaches task-specific model performance. Beyond imaging, Med-Gemini-Polygenic outperforms the standard linear polygenic risk score-based approach for disease risk prediction and generalizes to genetically correlated diseases for which it has never been trained. Although further development and evaluation are necessary in the safety-critical medical domain, our results highlight the potential of Med-Gemini across a wide range of medical tasks.

arXiv:2405.03162v1 [cs.CV] 6 May 2024

[†] See Contributions and Acknowledgments section for full author list.

Corresponding authors: {dangolden, shekazizi, kellych, roryp}@google.com.

1. Introduction

Medical data from diverse sources like biobanks, electronic health records, medical imaging, wearables, biosensors, and genomic sequencing are enabling the development of multimodal AI solutions that can better capture the complexity of human health and disease (Acosta et al., 2022). While AI in medicine has primarily focused on narrow tasks with single input and output types (Rajpurkar et al., 2022), recent advances in generative AI show promise in addressing multimodal, multi-task challenges in medical settings (Moor et al., 2023a,b).

The emergence of large language models (LLMs) and large multimodal models (LMMs) such as Flamingo (Alayrac et al., 2022), PaLI (Chen et al., 2022), GPT-4 (Achiam et al., 2023), GPT-4v (OpenAI, 2023), PaLM (Anil et al., 2023; Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), LLaVa (Liu et al., 2023, 2024a), and Mistral 7B (Jiang et al., 2023) that promise significantly enhanced context length and improved multimodal capabilities suggests that the realization of highly complex multimodal reasoning across various medical data will soon be achievable. These advancements have catalyzed the expansion of LLMs specifically designed for medical domains, such as Med-PaLM and its successor Med-PaLM 2 (Singhal et al., 2023a,b), Clinical Camel (Toma et al., 2023), MedAlpaca (Han et al., 2023), BioMistral (Labrak et al., 2024), sc-GPT (Cui et al., 2024), and others. Going beyond text alone, recent works have extended the capabilities of these base multimodal models by building models that cover various medical imaging modalities like Med-PaLM M (Tu et al., 2024), Med-Flamingo (Moor et al., 2023b) as well as those that focus on a specific imaging domain, such as radiology (Hamamci et al., 2024; Hyland et al., 2023; Tanno et al., 2024; Thawkar et al., 2023; Xu et al., 2023) and histopathology (Ikezogwo et al., 2024; Lu et al., 2024; Sun et al., 2024).

The release of the Gemini models (Gemini Team, Google, 2023; Google, 2024), with their advanced multimodal capabilities and breakthroughs in long-context understanding, marked a significant step forward in multimodal reasoning. Given its inherent human focus, medicine is a field in which advanced multimodal systems like Gemini are expected to be transformative (Acosta et al., 2022). Evaluations have already started to evaluate the base performance of these newer multimodal models (Pal and Sankarasubbu, 2024). However, the true potential of multimodal foundation models in the medical field remains largely underexplored due to the complexity of optimizing for problems in this field (Moor et al., 2023a; Rajpurkar et al., 2022) and a lack of diverse and meaningful evaluations that are grounded in clinical use cases (Fleming et al., 2023; Royer et al., 2024; Zhang et al., 2023a). To better understand the nuances of model capabilities and limitations, it is necessary to optimize multimodal models for a diversity of relevant clinical applications and rigorously evaluate them on appropriate clinical datasets.

This report details our efforts in exploring Gemini’s capabilities across a range of challenging multimodal medical tasks. Our evaluation benchmarks include 2D and 3D radiology images, histopathology patches, ophthalmology images, dermatology images, and genetic risk scoring. Our benchmark suite includes both open benchmark datasets and our own curated datasets. Open benchmark datasets have the advantage of being established and enabling direct comparison to others’ work, but they are often limited or methodologically flawed, leading to results that can overstate performance. For the custom benchmarks that we introduce, we have prioritized high quality metrics that are closely correlated to clinical utility. In particular, we focused on expert human evaluations for quantifying performance on CXR and CT report generation and on open visual question answering (VQA) questions from VQA-Rad. Additionally, we compared Med-Gemini to previous work or to the non-medically tuned version of Gemini where possible.

Where we believed it was helpful, we have proactively improved the quality of certain open benchmarks. This includes updating and correcting erroneous labels (such as MIMIC-CXR-JPG classification labels), extending the task scope of datasets (such as introducing VQA question/answer pairs for MIMIC-CXR), and refining data splits to remove train-test contamination (such as PAD-UFES-20 and VQA-Rad). We hope to release these improvements publicly soon.

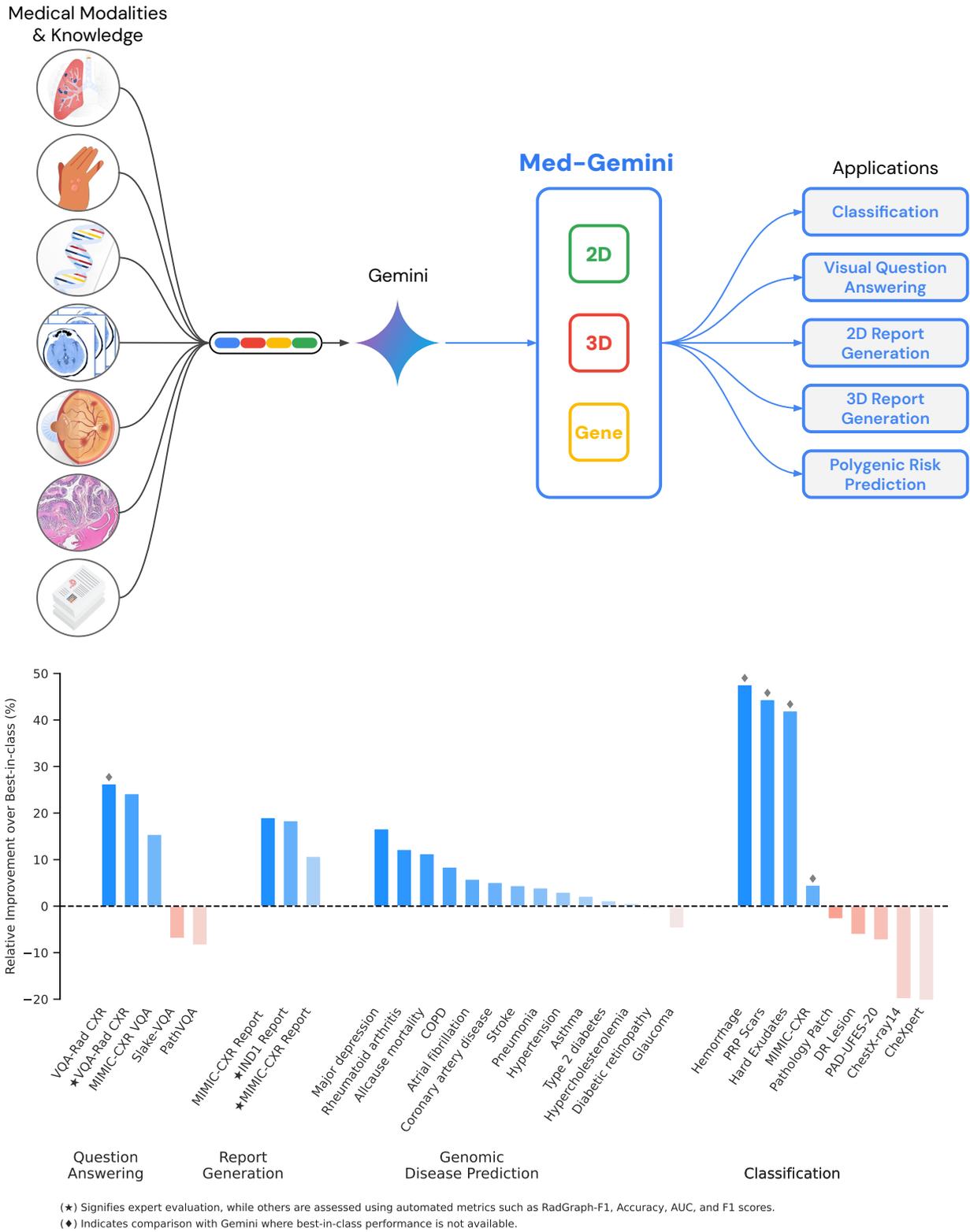


Figure 1 | Overview of our approach to curate and assess our family of medically tuned Gemini models, Med-Gemini. (top) These models build upon Gemini’s powerful capabilities in advanced reasoning, multimodal understanding, and long-context processing enriched with patient representation and medical knowledge. (bottom) Relative performance of Med-Gemini compared to SoTA or baselines across various tasks as detailed in Table A.15. Expert evaluation confirms that Med-Gemini-2D sets a new standard for AI-powered chest X-ray report generation, with relative improvements of 10% and 18% over the previous leading model across two distinct datasets. In histopathology, ophthalmology, and dermatology image classification, it surpasses baseline on 18 out of 20 tasks and approaches task-specific model performance. Med-Gemini-Polygenic outperforms the standard approach for disease risk prediction and generalizes to diseases for which it has never been trained.

In this report, we expand the fine-tuned family of models, Med-Gemini, specifically focusing on medical imaging and genomics. The models described here were tuned on a dataset of 7 million samples obtained from 3.7 million medical images and cases, spanning medical image classification, VQA, report generation, and genomic risk prediction, detailed in Section 2. Importantly, this dataset includes mostly free text paired with medical data, which eliminates the need for expensive expert labeling of the training data. We intentionally explore both medical image-based tasks and also the non-image-based task of polygenic risk prediction in order to evaluate the potential of Med-Gemini beyond imaging and in the crucial medical domain of long term risk prediction. Our findings demonstrate that LMMs have significantly advanced over the past year and are able to perform an increasing range of challenging tasks. Our key contributions are summarized as follows:

- **Med-Gemini:** A family of generalist medical AI models fine-tuned from Gemini, capable of performing a diverse set of medical tasks including medical image classification, VQA, report generation, and genomic risk prediction. Med-Gemini extends Gemini’s capabilities to include interpretation of diverse medical data, including both genomics and 2D and 3D medical images. Additional capabilities of Med-Gemini are described in “Capabilities of Gemini Models in Medicine” by [Saab et al. \(2024\)](#).
- **Clinically-relevant benchmarking:** We evaluate Gemini and Med-Gemini on a comprehensive set of clinically relevant benchmarks including 22 datasets across five different tasks and six distinct medical image modalities. Our evaluation suite includes eight out-of-distribution datasets to assess generalization capabilities of this new family of models. Our assessments primarily consist of automated metrics but we rely on expert human evaluation for tasks where expert human judgment is critical, namely chest X-ray and computed tomography (CT) report generation and radiology VQA on open questions in the VQA-Rad dataset.
- **Promising or best-in-class performance in several clinically relevant tasks:** Med-Gemini demonstrates best in class performance on chest X-ray and CT report generation and chest X-ray classification. Med-Gemini can also be used to predict disease and mortality risk more accurately than a standard linear polygenic risk score (PRS) based approach. Med-Gemini approaches the performance of models trained using orders of magnitude more training examples on dermatology, histopathology, and ophthalmology image classification and demonstrates competitive performance across several VQA tasks across pathology and radiology.

2. Datasets

Many different public and private datasets were used in the training and evaluation of Med-Gemini. All datasets were de-identified. Open datasets were used in accordance with their existing licenses and private datasets were used with permission and appropriate licenses.

2.1. Datasets for fine-tuning and instruction-tuning

Datasets were split into train, validation, and test sets by patient identifier when available. When patient identifiers were not available for a dataset, we ensured that there was no case or image overlap between splits.

2.1.1. Public datasets

MIMIC-CXR: MIMIC-CXR contains 377,110 images from 65,379 patients, with de-identified free-text reports describing the images ([Goldberger et al., 2000](#); [Johnson et al., 2019a,c](#)). This dataset is the largest public chest X-ray dataset, acquired in the emergency department of Beth Israel Deaconess Medical Center in the US. For each patient, there are multiple views and a corresponding report labeled for 13 common radiological conditions using the CheXpert labeler ([Irvin et al., 2019](#)) or with

“no finding” if no condition is present. Available labels include atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomeastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices, and no finding. We used the MIMIC-CXR training set (237,912 images) to fine-tune Gemini as described in Section 3 and detailed in Table 1. We further employed the test cases of MIMIC-CXR as a benchmark for multiple evaluation tasks including classification, report generation and VQA. For the report generation task, we used the chest X-ray image corresponding to the frontal view (anterior-posterior or posterior-anterior) to generate the *Findings* and *Impression* sections, similar to prior works (Tanno et al., 2024). For cases where no frontal view was available, we excluded them from our evaluation. For the VQA task, we utilized the condition-dependent VQA dataset (e.g. pleural effusion presence/location/severity) introduced in Xu et al. (2023), which we will make publicly available soon. In addition, we have used radiologist-adjudicated updated labels for findings that are also planned to be released soon (Park et al., 2024).

Mendeley digital knee X-ray images: This dataset consists of digital X-ray images of knee joints which were collected from hospitals and diagnostic centers. Original images are 8-bit grayscale. Each knee X-ray image was manually annotated by two medical experts following the Kellgren and Lawrence system for classification of osteoarthritis (Gornale and Patravali, 2020). There are a total of 1,633 unique images in this dataset, and we utilized 1,469 images in our training.

PAD-UFES-20: The dataset consists of 1,373 patients, 1,641 skin lesions, and 2,298 images. Skin lesion images were collected from various smartphones and exhibit variations in resolution, size, and lighting conditions. The dataset was acquired in collaboration with the Dermatological and Surgical Assistance Program (PAD) at the Federal University of Esp rito Santo (UFES-Brazil) (Pacheco et al., 2020). PAD is a non-profit program offering free skin lesion treatment, particularly to those who cannot afford private care. The dataset includes images of six different skin lesion types and diagnostics: three skin diseases and three skin cancers. These include basal cell carcinoma (BCC), squamous cell carcinoma (SCC), actinic keratosis (ACK), seborrheic keratosis (SEK), Bowen’s disease (BOD), melanoma (MEL), and nevus (NEV). We randomly split the dataset into train (corresponding to 90% of total samples) and test (10% of total samples). We utilized 2,047 training samples from this dataset in our training corpus when fine-tuning Gemini as described in Section 3 and detailed in Table 1. We intend to publicly release our dataset split soon.

National Lung Screening Trial (NLST): We utilized CT volumes from the validation subset of the NLST dataset (NLST, 2014) and processed them through a lung cancer screening system (Kiralay et al., 2024) to create a dataset of the most salient 2D slices from CT volumes. Captions were assigned based on the *scr_group* label in the NLST participant dictionary: values of 1 and 2 were considered as having nodules whereas a value of 3 were considered as not having a nodule. We then selected a total of 2,199 slices consisting of 1,324 studies with nodules and 875 without nodules for further analysis from the previously defined validation set (Ardila et al., 2019). The lung cancer screening system generated captions for each slice based on the first detected region. Slices without nodules received simpler descriptions, for example: “An axial CT slice of the middle lungs with no nodules.” For slices containing nodules, the captions included: location of the nodule (left or right lung), suspicion level for malignancy, and estimated size in millimeters based on the screening system’s output. We then split the slices into training and validation evenly based on the presence of nodules, allocating 80% for training, including 1,759 image and caption pairs, and 20% for validation, including 440 image and caption pairs. All 2D slice images were set to a window of $[-1400HU, 100HU]$.

Slake-VQA: This dataset is a large bilingual (English and Chinese) VQA dataset meticulously annotated by experienced physicians (Liu et al., 2021). It offers 642 images with 14,028 question-answer pairs in three imaging modalities (i.e. CXR, CT, MRI). Slake-VQA includes various areas of radiology, covering human body regions like the brain, neck, chest, abdomen, and pelvic cavity. The dataset comprises 9,849 VQA samples for training, 2,109 for validation, and 2,070 for testing. Questions are diverse,

including both open-ended (free-form) and closed-ended (yes/no) formats. They probe various image aspects such as plane, quality, position, organ, abnormality, size, color, shape, and related medical knowledge. We used only English-language examples from the official splits, which included 4,919 training, 1,053 validation, and 1,061 test examples.

PathVQA: This is a dataset of question-answer pairs on pathology images (He et al., 2020). The dataset includes both open-ended questions and closed-ended (yes/no) questions and is built with automated methods using two publicly-available pathology textbooks and a publicly-available digital library. The dataset includes 32,632 question-answer pairs on 4,289 images. The official training, validation, and test splits contain 19,654, 6,259, and 6,719 QA pairs. We leveraged the official train and test sets for training our model and evaluating its performance, respectively.

VQA-Med: The VQA-Med-2019 dataset offers a collection of medical images and associated question-answer (QA) pairs for model training and evaluation (Ben Abacha et al., 2021). It includes a training set with 3,200 medical images and 12,792 QA pairs, a validation set with 500 medical images and 2,000 QA pairs, and a test set containing 500 medical images and 500 questions. For the purpose of this report, we removed all images that overlapped with VQA-Rad (Lau et al., 2018) to avoid contamination. This resulted in 12,664 QA pairs used in training.

UK Biobank: Genetic factors play a significant role in an individual’s risk of developing various diseases. In this work, we used UK Biobank (Bycroft et al., 2018), a resource of nearly 500,000 de-identified individuals with genetic, lifestyle, and health information, to develop a task that takes as input an embedding of an individual’s genomic data and uses it to predict an individual’s status for various broad health outcomes. We extracted a set of 432,090 samples of European genetically inferred ancestry with genomic data passing quality control thresholds and split it randomly into train, validation, and test splits containing 60%, 20%, and 20% of the samples, respectively. Following best practices for polygenic risk prediction, we avoided including individuals who were genetically similar in two different data splits (Choi et al., 2020).

PMC-OA: PMC-OA is a medical dataset with image-caption pairs collected from PubMedCentral’s OpenAccess subset. Using the method described in Zhang et al. (2023b), we retrieved 3,110,109 scientific papers containing 15,505,259 image-caption pairs. To ensure meaningful analysis, we filtered for image-caption pairs containing at least one photographic image (e.g., excluding images corresponding to data figures), resulting in a final dataset of 2,246,656 image-caption pairs.

2.1.2. Private datasets

Histopathology patches: Pathological examination of tissue samples is crucial for effective diagnosis and treatment planning. Data from nine tasks across six tissue types from prior work (Lai et al., 2023) were used in our training set (Table A.8). Multi-class annotation masks were used for both sampling image patches from whole-slide images as well as generating captions. Patches of size 256×256 were sampled from whole slide images in a class-balanced manner. For each of the nine tasks, up to 10,000 image patches were sampled for three different magnification levels (2, 1, and 0.5 microns-per-pixel), resulting in 207,603 unique patches. Patch-level captions were created via prompting of a large language model (Gemini Pro) with inputs including structured slide-level metadata as well as patch-level annotation labels. Multiple captions per class for each task were generated and then manually reviewed to ensure an appropriate level of detail and accuracy, resulting in 5–7 captions per class across tasks. Combining the sampled patches with our curated captions resulted in 1,550,976 image-text pairs for fine-tuning. For examples of some of our curated captions corresponding to the annotation labels, see Table A.7.

Fundus images (EyePACS): Diabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. We used the de-identified dataset from EyePACS Inc. (Cuadros and Bresnick, 2009) and converted diabetic lesion-level presence labels to captions. The lesions considered were microaneurysms, hemorrhages, hard exudates, panretinal photocoagulation (PRP) scars, neovascularization of the disc and neovascularization elsewhere. For caption conversion, if a given image has lesion presence, for example, microaneurysm and hemorrhage, the associated generated caption was “microaneurysm is present, hemorrhage is present.” For healthy eyes, we use “no diabetic retinopathy related lesion” as the caption. 12,976 images with lesions and 3,000 healthy eye images were used to construct the dataset.

Computed tomography images (CT-US1): A comprehensive dataset comprising 753,247 CT studies with associated radiology reports from 615,384 patients was obtained from three major hospital regions in the United States. These CT studies included head/neck, chest, heart, abdominal, spine, and extremity regions imaged with and without contrast. To ensure robust evaluation, we employed a patient-level random split for training, validation, and testing. The data was divided into 70% for training, 15% for validation, and 15% for testing on the patient level. After an ingestion process this resulted in 657,719 training volumes and a total of 23,649 validation volumes. Due to the reliance on expert evaluation for report generation, a subset of 92 non-contrast head/neck CT volumes from unique patients in the test set was used for model assessment. Volumes were prepared as described in Section 2.3. Only axial image volumes containing more than 10 slices were included in the prepared data and the volume within the study with the most axial slices was selected for inference.

We also carefully processed the existing dataset to create an extra 2D CT slice dataset specifically tailored for training our 2D model. This involved filtering radiology reports for specific series and image numbers, selecting the correct images, and windowing them to a window of $[-1000HU, 100HU]$. To ensure that the text pertained directly to the CT slice in question and was a comprehensive description of it, captions were generated by combining the sentence of the report referencing the image along with the following sentence. This process resulted in a dataset of 4,009 images consisting of 3,207 training and 802 validation examples, primarily focused on CT studies of the abdomen and pelvis.

Chest X-ray images (CXR-US2): The CXR-US2 dataset corresponds to the training set of US1 in Xu et al. (2023). This dataset consists of 132,680 frontal chest X-ray images from 12,988 patients taken at an academic medical center in Illinois, USA. Further descriptive statistics can be found in Xu et al. (2023).

2.2. Held-out datasets for evaluation and benchmarking

Beyond the test sets associated with our training datasets (described above), we also utilized multiple held-out and out-of-distribution (OOD) datasets.

2.2.1. Public datasets

CheXpert The CheXpert dataset is similar to the MIMIC-CXR dataset and consists of 224,316 chest X-ray images (both frontal and lateral views) from 65,240 patients (Irvin et al., 2019). It labels 14 distinct thoracic conditions, including “No Finding”. The original CheXpert dataset contains positive, negative, uncertain and unmentioned labels. During evaluation we considered the “unmentioned” label as negative and included only chest X-rays depicting frontal views.

VQA-Rad The VQA-Rad dataset (Lau et al., 2018) comprises 315 radiology images sourced from CT, MRI, and X-ray scans, and it encompasses three anatomical regions including the head, abdomen, and chest. This dataset includes a wide array of question types, spanning 11 distinct categories, such as modality, plane, organ system, abnormality, and more, where 58% of the question-answer pairs are designed to be closed-ended (yes/no or limited choices), while the remaining 42% are open-ended.

Table 1 | Overview of the training datasets. More than 7 million data samples from 3.7 million medical images and cases is used for fine-tuning and further instruction-tuning of Gemini for medical applications in Med-Gemini. This includes diverse set of modalities including 2D and 3D radiology images, pathology, ophthalmology, and genomic data. These datasets includes mostly free text paired with medical data, which eliminates the need for expensive expert labeling of the training data.

| Modality | Dataset | No. examples | No. Images | Description |
|----------------|--------------------|--------------|------------|---|
| Radiology (2D) | Slake-VQA | 4,919 | 450 | Radiology images & QA pairs |
| | MIMIC-CXR | 2,142,892 | 231,483 | Radiology images & free-form reports |
| | Digital Knee X-ray | 1,469 | 1,469 | Knee X-ray images & labels |
| | CXR-US2 | 132,680 | 132,680 | Radiology images & free-form reports |
| | NLST | 2,199 | 2,199 | 2D CT slices & free-form reports |
| | CT-US1 | 3,207 | 3,207 | 2D CT slices & free-form reports |
| Radiology (3D) | CT-US1 | 657,719 | 657,719 | 3D CT images & free-form reports |
| Pathology | PathVQA | 19,654 | 2,599 | Pathology images & QA pairs |
| | Histopathology | 1,550,976 | 207,603 | Histopathology images, captions, & QA pairs |
| Dermatology | PAD-UFES-20 | 2,047 | 2,047 | Skin lesion images & labels |
| Ophthalmology | EyePACS | 14,406 | 14,406 | Fundus images & labels |
| Medical VQA | PMC | 2,246,656 | 2,246,656 | PubMed Central images & caption pairs |
| | MedVQA | 12,664 | 3,168 | Medical images & QA pairs |
| Genomics | UK Biobank | 259,225 | 259,225 | Genomic data & disease outcomes |

The standard and official splits of the dataset feature 1,797 QA pairs for training and 451 for testing purposes. However, due to contamination of images included in both training/test in the original dataset release, we constructed a new, non-overlapping test and tuning split for the subset of chest X-ray images and associated question-answer pairs, first described in [Xu et al. \(2023\)](#). In this study, we went one step further and created new image-disjoint splits of train, validation and test sets for all three image types. We ensured that the previous X-ray-only validation and test sets were subsets of the new validation and test sets, respectively, thus enabling comparisons with the ELIXR model ([Xu et al., 2023](#)) on the new test set, as no former validation examples were included in the new test set. Aside from these constraints, we sampled in a manner that roughly equalizes both the ratio of closed to open question-answer pairs for each anatomical region, see [Table A.5](#), as well as the distribution across the 11 different question types across the three splits, see [Table A.6](#) in [Section A.1.3](#). Henceforth, we refer to this new three-way split as the “balanced split”. In total, the balanced VQA-RAD dataset split, which we will make publicly available soon, comprises 2,248 pairs of questions and answers, encompassing 1,299 closed-ended questions and 949 open-ended questions.

ChestX-ray14 The ChestX-ray14 dataset ([Summers, 2019](#)) is a comprehensive medical imaging dataset containing 112,120 frontal-view chest X-ray images from 30,805 unique patients. ChestX-ray14 builds upon the ChestX-ray8 dataset ([Wang et al., 2017](#)), expanding the number of labeled diseases to fourteen common thoracic pathologies, including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule, Mass, and Hernia. Because these labels are automatically derived using NLP techniques and therefore contain inherent uncertainty, we restricted our evaluation to a subset of 1,962 cases focusing on three radiologist-adjudicated conditions ([Majkowska et al., 2020](#)), namely lung opacity, pneumothorax, and fracture.

TCGA study type This dataset utilizes histopathology images from The Cancer Genome Atlas (TCGA), for which different study types correspond to different cancer types with additional information via portal.gdc.cancer.gov. The patches from this dataset are sampled from 2,952 training slides, 1,466 validation slides, and 1,489 test slides across ten (10) distinct TCGA study types: BLCA, BRCA, COAD,

HNSC, KIRC, LIHC, LUAD, LUSC, OV, and STAD (Lai et al., 2023). We used the test set as an out-of-distribution dataset to evaluate our model’s ability to generalize to different histopathology-related tasks.

2.2.2. Private datasets

IND1 This is a private research dataset of a similar scale as MIMIC-CXR, which we refer to as IND1 (Nabulsi et al., 2021). This dataset comprises 263,021 de-identified frontal chest X-rays (digital and scanned) along with their corresponding reports. The X-rays were collected from five regional centers (Bangalore, Bhubaneswar, Chennai, Hyderabad, and New Delhi) across a large hospital group in India between November 2010 and January 2018 (Ahn et al., 2022). We used the same test set as (Tanno et al., 2024), and following their framework, 300 of those cases are used for human evaluation.

TTH tissue type The TTH tissue type dataset, introduced by Weng et al. (2019) and Lai et al. (2023), represents a patch-level tissue type classification task. This internal dataset comprises 17,319 training slides, 6,488 validation slides, and 6,719 test slides, encompassing a total of 16 distinct tissue types. These tissue types include Appendix, Breast, Cervix, Colon and Rectum, Fallopian Tube, Gallbladder, Liver, Lymph Node, Ovary, Placenta, Prostate, Skin, Thyroid, Upper GI, Uterus, and Vas Deferens. We used the test set to evaluate generalization of our model to different histopathology tasks.

2.3. Data preprocessing

Radiology 2D images When available, images acquired in DICOM format were used to directly create examples for training and inference. In the case of X-rays, raw pixel data were extracted from the DICOM image pixel data, and the look up table (LUT, part of the DICOM metadata) was subsequently applied. If a DICOM file contained multiple LUTs, we used the first LUT entry. If the window width and window center were defined, these were also used for preprocessing. The final pixel data were re-scaled to the full range of [0, 65535] for the 16-bit PNG format. X-ray images in a preprocessed format were taken as is.

CT volumes All 3D CT volumes were derived from DICOM images. Only axial slices were used to establish a standardized anatomical perspective. Slices were sorted based on the Image Position (Patient) attribute and used to compute slice spacing and reconstruct volumes. Subsequently, images were clipped with a Hounsfield Unit (HU) range of $[-1024HU, 1024HU]$ to cover a full spectrum of densities (e.g. the typical window/level values of brain, soft tissues) and then scaled to [0.0, 1.0]. Finally, tricubic interpolation was used to resample all images to a voxel spacing of $0.7\text{mm} \times 0.7\text{mm} \times 1.4\text{mm}$, ensuring consistent uniform resolution for accurate comparative analysis.

Genomics: A genomic featurization for an individual consists of polygenic risk scores (PRSs) for 7,415 traits. Each PRS estimates the genetic risk of the individual for a particular disease or trait, calculated by aggregating the estimated effects of many common variants associated with the condition. Each PRS was computed using genome-wide association study summary statistics computed by the Pan-UKB Consortium (Pan-UKB team, 2020). These genomic features were then converted to images by projecting the PRSs into patch-aligned squares of 8×8 pixels with values between [0, 255]. The 3 RGB channels of the images were used to stack 3 different p-value thresholds of the projections. The PRS features were obtained from the genetic information of 314,540 individuals of European genetically inferred ancestry from the UK Biobank (Bycroft et al., 2018; Sudlow et al., 2015).

To create training and evaluation labels, we selected eight in-distribution health outcomes which have strong heritability (i.e. genetic information plays an important role in influencing susceptibility (Visscher et al., 2008)), span multiple organ systems, and are challenging to predict from polygenic

Table 2 | Overview of the datasets used for evaluating our fine-tuned Gemini models, Med-Gemini. Our evaluation leveraged a robust dataset suite encompassing 22 datasets across 5 different types of clinically relevant tasks. This included 8 out-of-distribution datasets to assess generalization and spanned 7 distinct medical image modalities. We explicitly explored medical image classification, VQA, 2D report generation, 3D report generation, and disease prediction from genetic risk embeddings. The total number of evaluation samples across these datasets exceeds 40,000.

| Task | Dataset/Setup | Modality | No. Samples | OOD |
|------------------------------------|---------------------------------------|----------------|-------------|-----|
| Radiology report generation (2D) | MIMIC-CXR | Chest X-ray | 912 | - |
| | MIMIC-CXR (Expert Evaluation) | Chest X-ray | 206 | - |
| | IND1 (Expert Evaluation) | Chest X-ray | 300 | ✓ |
| Radiology report generation (3D) | CT-US1 (Expert Evaluation) | CT | 92 | - |
| VQA | MIMIC-CXR VQA | Radiology | 226 | - |
| | Slake-VQA (English-only) | Radiology | 1,061 | - |
| | VQA-Rad | Radiology | 2,248 | ✓ |
| | PathVQA | Pathology | 6,719 | - |
| Classification | PAD-UFES-20 | Dermatology | 251 | - |
| | MIMIC-CXR (Abnormal/Normal) | Chest X-ray | 2,242 | - |
| | MIMIC-CXR (5 classes) | Chest X-ray | 2,242 | - |
| | CheXpert (Abnormal/Normal) | Chest X-ray | 1,962 | ✓ |
| | CheXpert (5 classes) | Chest X-ray | 518 | ✓ |
| | ChestX-ray14 (3 classes) | Chest X-ray | 1,962 | ✓ |
| | CAMELYON16 (2 classes) | Histopathology | 258 | - |
| | Gleason NCB (4 classes) | Histopathology | 88 | - |
| | Gleason RP (4 classes) | Histopathology | 202 | - |
| | Lung AD (9 classes) | Histopathology | 202 | - |
| | Breast IC (3 classes) | Histopathology | 669 | - |
| | Breast NP (3 classes) | Histopathology | 945 | - |
| | Breast TF (3 classes) | Histopathology | 945 | - |
| | CIN (3 classes) | Histopathology | 229 | - |
| | CRC (2 classes) | Histopathology | 44 | - |
| | TCGA Study Type (10 classes) | Histopathology | 1,489 | ✓ |
| | Tissue Type (16 classes) | Histopathology | 6,719 | ✓ |
| | EyePACS Hard Exudates (2 classes) | Ophthalmology | 498 | - |
| EyePACS Hemorrhage (2 classes) | Ophthalmology | 498 | - | |
| EyePACS DR Lesion(Abnormal/Normal) | Ophthalmology | 490 | - | |
| Risk Prediction (UK Biobank) | Coronary artery disease | Genomic | 400 | - |
| | Stroke | Genomic | 400 | - |
| | Type 2 diabetes | Genomic | 400 | - |
| | Glaucoma | Genomic | 400 | - |
| | Chronic obstructive pulmonary disease | Genomic | 400 | - |
| | Rheumatoid arthritis | Genomic | 400 | - |
| | Major depression | Genomic | 400 | - |
| | All-cause mortality | Genomic | 400 | - |
| | Hypertension | Genomic | 400 | ✓ |
| | Hypercholesterolemia | Genomic | 400 | ✓ |
| | Atrial fibrillation | Genomic | 400 | ✓ |
| | Diabetic retinopathy | Genomic | 400 | ✓ |
| | Pneumonia | Genomic | 400 | ✓ |
| | Asthma | Genomic | 400 | ✓ |

risk scores alone: coronary artery disease, stroke, type 2 diabetes, glaucoma, chronic obstructive pulmonary disease (COPD), rheumatoid arthritis, major depression, and all-cause mortality. Additionally, to assess model generalization, we selected six out-of-distribution (OOD) health outcomes that share genetic correlation with one or more of the in-distribution health outcomes: hypertension, hypercholesterolemia, atrial fibrillation, diabetic retinopathy, asthma, and pneumonia (Table A.10).

Pathology patches Patches with initial size of 256×256 pixels were sampled from whole slide images using multi-class annotation masks in a class-balanced manner across three different magnification levels (2, 1, and 0.5 microns-per-pixel).

Input preprocessing and tokenization Images from all 2D datasets were uniformly resized to 768×768 pixels, preserving aspect ratio with padding, with pixel intensities scaled to $[0, 1]$. This ensured image resolution would be high enough for the fine-grained detail of medical images. For text, we used the native Gemini SentencePiece tokenizer (Gemini Team, Google, 2023; Kudo and Richardson, 2018) without modification.

3. Modeling Methodology

3.1. Model architecture

Gemini builds upon the robust foundation of Transformer decoders (Parmar et al., 2018; Vaswani et al., 2017), offering significant architectural and optimization enhancements for efficient, stable large-scale training (Barham et al., 2022; Gemini Team, Google, 2023). This equips Gemini with exceptional natural language understanding and text generation capabilities. Of particular interest for medical data processing, Gemini’s multimodal design draws inspiration from foundational Google research on Flamingo (Alayrac et al., 2022), CoCa (Yan et al., 2022; Yu et al., 2022) and PaLI (Chen et al., 2022), enabling enhanced multimodal understanding and reasoning.

Gemini handles video understanding by encoding frames as a sequence within its large context window (Gemini Team, Google, 2023; Google, 2024). This allows seamless integration of video frames, multi-slice images, text, or audio inputs. The model even supports variable input resolutions, enabling it to prioritize computational resources for tasks requiring high-resolution analysis. Gemini 1.5 specifically is a mid-size model with a context window of up to 1 million tokens and performance on par with the largest Gemini model, 1.0 Ultra. Given this exceptional efficiency, we chose to finetune Med-Gemini from Gemini 1.5.

3.2. Multimodal fine-tuning

Three custom versions of the Gemini 1.5 Pro vision encoder were trained for 2D modalities, 3D modalities, and genomics. In our initial experiments, we found that custom vision encoders for each type of data format performed better than a single vision encoder for all data formats. Furthermore, fine-tuning the vision encoder as well as the language component in Gemini led to significantly better visual understanding in comparison to a model that used the native vision encoder of Gemini 1.5 Pro models. From these three custom vision encoders, we trained three specific variants of Med-Gemini which we refer to as Med-Gemini-2D, Med-Gemini-3D, and Med-Gemini-Polygenic. Notably, Med-Gemini-2D includes all conventional medical images that are encoded in 2D (e.g. chest X-ray, CT slices, pathology patches), Med-Gemini-3D was built on top of Med-Gemini-2D and handles 3D medical data (e.g. CT), and Med-Gemini-Polygenic was trained for a novel image encoding derived from non-image features (e.g. genomics). For all three model variants, fine-tuning was framed as a captioning or VQA task.

Fine-tuning for 2D modalities - Med-Gemini-2D All 2D modalities were fine-tuned together using the training mix described in Section 2 and Table 1 to create Med-Gemini-2D. The 2D modalities used for fine-tuning included the described radiology, pathology, dermatology, and ophthalmology images.

Fine-tuning for 3D modalities - Med-Gemini-3D To interpret 3D medical data, we leveraged the video understanding capabilities of Gemini (Google, 2024). Use of the Gemini video encoder allows Med-Gemini-3D to process multiple 2D slices, replacing the time axis with the depth dimension, with computed tomography (CT) as our example modality. This 3D fine-tuned model can then synthesize information across a series of 2D slices to generate radiology reports. Use of this video encoding capability will permit analysis of other volumetric and time-series medical data (e.g. MRI, ultrasound) in the future.

Fine-tuning for genomics - Med-Gemini-Polygenic Genomics “images” (polygenic risk scores (PRS) projected into 2D, see Section 2.3) were included in the mixture of datasets used to fine-tune the Med-Gemini-Polygenic vision encoder, and were trained to predict eight broad health outcomes (coronary artery disease, stroke, type 2 diabetes, glaucoma, chronic obstructive pulmonary disease, rheumatoid arthritis, major depression, and all-cause mortality) in a captioning task.

Instruction fine-tuning To optimize the instruction-following capabilities of the fine-tuned Med-Gemini even further, we subsequently employed an instruction-tuning phase. In this phase, we fine-tuned Gemini 1.5 Pro on a curated collection of multimodal data consisting of carefully crafted instruction and response pairs. By exposing the model to these examples, we refined its ability to not only understand the content of medical images and signals, but also to follow nuanced instructions and generate tailored outputs.

3.3. Model training and inference infrastructure

Like its predecessor Gemini 1.5 Pro and all other Gemini models, Med-Gemini was trained on large-scale Google TPUV4 accelerator pods spread across multiple data-centers. This training setup significantly scales up from our previous flagship PaLM family (Chowdhery et al., 2023). The Gemini architecture ensures efficient serving on TPU accelerators at scale. For detailed information on training and serving Gemini models, see (Gemini Team, Google, 2023; Google, 2024).

4. Evaluation and Results

The following sections explore in detail how Gemini and Med-Gemini perform across various modalities and tasks in the medical field. Due to restrictions in our data licenses, our evaluation was limited to internal models. Our evaluation leveraged a robust dataset suite encompassing 22 datasets across four different clinically relevant tasks (report generation, VQA, classification, risk prediction). This dataset includes eight out-of-distribution datasets to assess generalization and spanned seven distinct medical image modalities. An overview of the evaluation datasets is provided in Table 2. The total number of evaluation samples across these datasets exceeded 40,000.

4.1. Medical image classification

To rigorously evaluate Med-Gemini’s in-distribution and out-of-distribution performance, we employed a comprehensive medical image classification benchmark. This benchmark encompassed diverse modalities: skin lesion classification, chest X-ray classification, histopathology patch classification, and fundus image classification. We approached classification as a generative multi-choice task for zero-shot classification (no supporting example in prompt) and linear probing for label-efficient setups. This design allowed for a thorough assessment of Med-Gemini’s robustness and adaptability

across various medical imaging domains.

Chest X-ray image classification Our chest X-ray image classification evaluation focused on two key classification scenarios. First, we considered multi-label classification for the presence of each of five types of frequently occurring conditions: atelectasis, cardiomegaly, consolidation, pulmonary edema, and pleural effusion. This follows the suggestions from [Azizi et al. \(2021, 2023\)](#); [Irvin et al. \(2019\)](#); [Tanno et al. \(2024\)](#). Second, we performed binary classification for all images as either normal or abnormal, based on the CheXpert “no finding” label for frontal chest X-rays ([Irvin et al., 2019](#)). These two scenarios are used consistently across MIMIC-CXR and our out-of-distribution dataset CheXpert ([Irvin et al., 2019](#)). In addition, for ChestX-ray14 ([Summers, 2019](#); [Wang et al., 2017](#)) we focused the evaluation of our model on three specific conditions including lung opacity, pneumothorax, and fracture. For all test tests, we only included images that were frontal view (i.e. view position “AP” or “PA”). In addition, for MIMIC-CXR it is required the original report to contain a “Findings” section that could be extracted via regular expression matching.

We manually explored the validation set, prompting each model either with a multi-select prompt for all labels (i.e., 5 conditions plus normal/abnormal) at once, or multiple binary Yes/No prompts for each label separately, and found binary prompts to yield better Macro F1 results for Med-Gemini for all labels, and slightly better results for Gemini Ultra except for predicting normal/abnormal. The prompts used for evaluation are listed in Section [A.1.2](#). Answers were generated using nucleus sampling with a temperature of 0.0, a top_p of 0.75 and an output token limit of 200. Generated answers were normalized and matched against “yes”/“no” ground truth strings, which directly corresponded to 1.0 and 0.0 label values for all evaluation data sets. For multi-label, multi-class scenarios, we evaluated the average accuracy using the class-weighted F1 score. Details of the metrics used can be found in Section [A.3](#). The MIMIC-CXR labels were revised based on a selective review of flagged reports by board-certified radiologists. See Section [A.1.1](#) for more details about the revised MIMIC-CXR labels. For the MIMIC CXR evaluations, we excluded case/condition combinations with an “uncertain” (-1.0) or no label (blank), except for the “No Findings” condition, where all cases with a non-positive (1.0) label were considered negative. Classification results using data-efficient learning are described separately in Section [A.2.1](#).

Table [3](#) shows the comparison of the performance on the chest X-ray classification task between Med-Gemini and Gemini Ultra for in- and out-of-distribution datasets. Our medically tuned model outperformed Gemini Ultra across most labels on the in-distribution MIMIC-CXR dataset. Notably, we demonstrated significantly stronger performance on the normal/abnormal classification despite using a multi-select prompt for Gemini, which specified all the “abnormal” conditions (and yielded better results on the validation set than a dedicated binary prompt), versus a very short normal/abnormal binary prompt for Med-Gemini. However, on the more challenging out-of-distribution datasets (CheXpert and ChestX-ray14), performance is varied. Med-Gemini excels in some tasks such as cardiomegaly or pleural effusion detection on CheXpert, while lagging in others like fracture detection in ChestX-ray14, which is a strong minority class there. These results suggest room for improvement in handling significant domain shifts.

Histopathology image classification We evaluated the image embeddings of Med-Gemini-2D via linear probing on the 11 tasks from [Lai et al. \(2023\)](#) and summarized in Table [A.8](#). Med-Gemini was fine-tuned on data corresponding to 9 of these tasks (in-distribution), while 2 tasks were held-out (out-of-distribution). Together, these evaluation tasks cover a total of 17 tissue types, 12 cancer types, and several different types of classification tasks (e.g., tumor identification, grading, subtyping) across 3 magnifications. Linear probing was done as in [Lai et al. \(2023\)](#): briefly, a logistic regression model with L2-regularization was fit for each task, and task-specific regularization weights and magnifications were selected using the validation sets. Linear probe metrics on the test sets were calculated using 5,000 patches with logistic regression models trained on embeddings from the 10,000 train set plus 5,000 validation set patches. Confidence intervals for macro-averaged AUCs were

Table 3 | Med-Gemini-2D Performance on the chest X-ray classification task. Comparison of the performance on the chest X-ray classification task between Med-Gemini, Gemini Ultra and also SoTA, if comparable or available. For MIMIC-CXR datasets we utilized the revised labels (see Section A.1.1) denoted by †. Our medically fine-tuned model demonstrated superior performance on the in-distribution MIMIC-CXR dataset. However, for out-of-distribution datasets, Med-Gemini excelled in specific tasks such as Cardiomegaly detection, but both Gemini Ultra and Med-Gemini fell short of models exclusively trained for chest X-ray classification, and for strong minority class with distinctly different visual features such as Fracture.

| Dataset | Condition | F1 score(%) | | | OOD |
|--------------|--------------------|-------------|--------------|---------------|-----|
| | | Med-Gemini | Gemini Ultra | SoTA | |
| CheXpert | Atelectasis | 49.7 | 51.1 | 64.6 ‡ | ✓ |
| | Cardiomegaly | 72.0 | 51.5 | 74.3 ‡ | |
| | Consolidation | 23.0 | 17.9 | 33.3 ‡ | |
| | Edema | 32.7 | 46.2 | 60.2 ‡ | |
| | Pleural Effusion | 64.4 | 46.6 | 70.4 ‡ | |
| | Macro-F1 | 48.3 | 42.6 | 60.6 ‡ | |
| | Normal/Abnormal | 54.3 | 52.7 | - | |
| ChestX-ray14 | Lung Opacity | 79.9 | 73.3 | 88.8 * | ✓ |
| | Pneumothorax | 55.3 | 29.3 | 58.4 * | |
| | Fracture | 5.5 | 0 | 27.8 * | |
| | Macro-F1 | 46.7 | 34.2 | 58.3 * | |
| MIMIC-CXR | Atelectasis † | 99.8 | 88.1 | | - |
| | Cardiomegaly † | 94.1 | 94.6 | | |
| | Consolidation † | 82.0 | 77.0 | § | |
| | Edema † | 86.8 | 86.4 | | |
| | Pleural Effusion † | 90.8 | 88.1 | | |
| | Macro-F1 † | 90.7 | 86.8 | | |
| | Normal/Abnormal | 42.0 | 29.7 | | |

† Revised labels (Park et al., 2024).

‡ Results from CheXzero model (Tiu et al., 2022).

§ SoTA not compatible with test set nor labels.

* Labels and prediction from Majkowska et al. (2020). F1 scores computed from their PPV and sensitivity.

computed via blocked bootstrap (blocking on slides) with 10,000 replicates. For comparison, we also report performance with embeddings from an ImageNet21k-based ViT-S/16 model trained using the AugReg method (Steiner et al., 2021), embeddings produced by the vision encoder in Gemini Ultra, and embeddings from a histopathology-specialized model trained via self-supervision (Lai et al., 2023) (PathSSL). Results are reported in Figure 2. While the PathSSL embedding model is specialized to the histopathology domain, the image embeddings in Med-Gemini-2D achieved comparable performance while also demonstrating strong results across multiple other clinical domains.

Skin lesion classification Med-Gemini-2D achieves competitive classification accuracy using just dermatological images alone as input, and does not rely on metadata (e.g. patient demographics, lesion symptoms, living conditions). Such metadata, while provided in PAD-UFES-20, are not always readily available in clinical settings. We note that our evaluation is not directly comparable to Med-PaLM M (Tu et al., 2024) since (a) Med-PaLM M inputs an additional 14 clinical attributes and (b) we created different train and test splits to remove patient overlap between splits in the original Med-PaLM M work (we hope to publicly release these updated splits soon). To establish context for model performance, we compared Gemini performance with Derm Foundation (Google, 2024) which is a specialized dermatology model developed by Google. We trained linear probing classifiers on top of the Derm Foundation embeddings in this comparison.

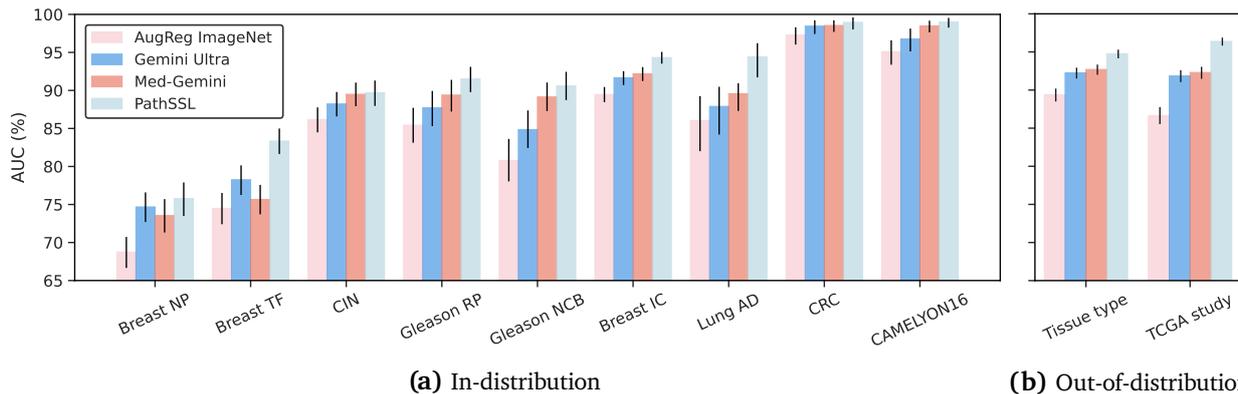


Figure 2 | Med-Gemini-2D histopathology image classification performance. Linear-probing on histopathology patch-classification tasks (macro-averaged AUC percentages with 95% confidence intervals) on our in-distribution and held-out out-of-distribution tasks. Overall, while Med-Gemini-2D outperforms Gemini Ultra on 7 out of 9 in-distribution tasks and on both out-of-distribution tasks, it does not improve over a histopathology-specific foundation model (PathSSL) on any of our tasks.

Table 4 | Performance on PAD-UFES-20 classification task. AUC linear probing demonstrates that Gemini Ultra and Med-Gemini-2D generate robust skin lesion classification embeddings, with Med-Gemini-2D further approaching the performance of the specialized Derm Foundation model.

| Metric | Weighted-AUC (%) | Weighted-F1 (%) | Accuracy (%) |
|--------------------------------|-----------------------------|-----------------------------|-----------------------------|
| Gemini Ultra | 92.6 (89.8, 95.1) | 60.3 (54.6, 66.9) | 60.6 (54.1, 66.8) |
| Med-Gemini-2D | 92.1 (89.2, 94.7) | 71.4 (65.4, 77.4) | 73.3 (67.7, 78.9) |
| Derm Foundation (Google, 2024) | 94.0 (91.6, 96.1) | 77.0 (71.4, 82.4) | 76.9 (71.3, 82.5) |

We utilized the following three metrics for evaluation. (1) Weighted-AUC (by class prevalence): We first extracted the embedding outputs from Med-Gemini-2D’s image encoder, Gemini Ultra’s image encoder and Derm Foundation. Then we individually trained linear probes on top of the embeddings on the entire PAD-UFES-20 train split to classify 6 skin lesion types, and computes their weighted-AUC on the test split. (2) Weighted-F1 (by class prevalence): For Med-Gemini-2D and Gemini Ultra, we extracted classification prediction based on string matching from the model output. For Derm Foundation, we took the argmax of the linear probe prediction as the predicted class. (3) Accuracy: We computed the classification prediction follows the same method as in weighted-F1.

Table 4 shows the performance comparison. From the AUC linear probing results, we can see that both Gemini Ultra and Med-Gemini-2D produced robust embeddings for skin lesion classification and had on-par performance when compared with the specialized Derm Foundation model. From F1 and accuracy, we can see that the fine-tuning in Med-Gemini-2D improved the LLM’s understanding of the embedding space for skin lesions and achieved performance close to the Derm Foundation model.

Fundus image classification We evaluated the performance of Med-Gemini-2D on four ophthalmology classification tasks. First, we approached the identification of three common Diabetic Retinopathy (DR) lesions, more specifically hard exudates, hemorrhage, and panretinal photocoagulation (PRP) scars, as a multi-label classification challenge. Then, we treated anomaly detection as a binary classification problem, where the model was tasked with determining the presence or absence of DR lesions in a fundus image. EyePACS dataset (Cuadros and Bresnick, 2009) was used for all tasks, with fundus images balanced for equal distribution of positive and negative labels.

Table 5 | Performance comparison of Med-Gemini-2D, Gemini Ultra, and a supervised model that is trained with additional data for fundus image classification. For all tasks, Med-Gemini-2D demonstrates significant improvement over Gemini Ultra. For lesion presence classification, Med-Gemini-2D is on-par with the supervised model for sensitivity, while specificity still have room for improvement (likely due to orders of magnitude less data used during the training process).

| Condition | Model | Accuracy (%) | Sensitivity (%) | Specificity (%) | F1 score (%) |
|--------------------|---|--------------|-----------------|-----------------|--------------|
| Hard exudates | Gemini Ultra | 55.2 | 71.5 | 39.0 | 61.5 |
| | Med-Gemini-2D | 88.4 | 80.3 | 96.4 | 87.3 |
| Hemorrhage | Gemini Ultra | 49.0 | 69.9 | 28.1 | 57.8 |
| | Med-Gemini-2D | 84.7 | 88.4 | 81.1 | 85.3 |
| PRP Scars | Gemini Ultra | 57.2 | 56.8 | 57.6 | 57.0 |
| | Med-Gemini-2D | 84.6 | 71.6 | 97.6 | 82.3 |
| DR lesions present | Gemini Ultra | 58.6 | 73.6 | 43.8 | 63.9 |
| | Med-Gemini-2D | 84.9 | 96.3 | 73.5 | 86.4 |
| | Supervised model trained with additional data | 94.7 | 96.8 | 92.6 | 94.9 |

We benchmarked Med-Gemini-2D against Gemini Ultra, which used both the fundus image and a multiple-choice-like prompt for prediction. To extract the prediction labels, we searched for specific markers that the LLM was constrained to output (e.g. (G) if no DR lesion is present). In contrast, Med-Gemini-2D relied only on the image, with prediction labels extracted by searching for keywords, like “hemorrhage.” For the anomaly detection task, we compare to a third model similar to (Krause et al., 2018) that has been trained using supervised learning to detect different grades of DR (none, mild, moderate, severe, proliferative). If the predicted DR grade is “none,” the fundus is considered normal (no DR lesion), otherwise, the image is predicted to have DR lesions present. We note that this supervised model was carefully trained on a much larger dataset containing more than 3 million fundus images from diverse manufactures/data sources/geography, and could be considered as an “upper bound” of this task.

Table 5 displays the performance of Med-Gemini-2D and Gemini Ultra on the classification of hard exudates, hemorrhages and PRP Scars, utilizing accuracy, sensitivity, specificity, and F1 score as evaluation metrics. It also compares the performance of DR lesions detection between, Med-Gemini-2D, Gemini Ultra and the strong supervised model. Results demonstrate that Med-Gemini-2D consistently outperformed Gemini Ultra on both multi-label and binary classification tasks. Notably, Med-Gemini-2D achieved significantly higher specificity in hard exudate classification (96.4% vs. Gemini Ultra’s 39.0%) and hemorrhage classification (81.1% vs. Gemini Ultra’s 28.1%). These results highlight the benefits of task-specific fine-tuning for this highly-specialized medical domain. Med-Gemini-2D underperformed on the anomaly detection task compared to the strong supervised model, but it is important to acknowledge the strong supervised model’s significant advantage in the $\sim 200\times$ volume of labeled data used during its training process.

For three other attempted classification tasks, namely detection of microaneurysms, neovascularization of the optic disc and neovascularization elsewhere, Med-Gemini-2D appeared to be miscalibrated, predicting most cases as negative in the LLM text output. While this is likely related to the training dataset distribution and overall data mixing ratio, further work is needed to improve question-answering based classification and calibration.

Table 6 | Evaluation details for VQA tasks. Our Med-Gemini-2D’s VQA performance was evaluated across diverse medical specialties using various datasets. Notably, in radiology tasks, Med-Gemini-2D outperformed Gemini and previous best-in-class performance across different subsets and metrics, improving best-in-class performance by over 11%. Performance is reported using Mean Tokenized F1-score, Mean Expert Score, and Accuracy. Based on radiologist assessments, we excluded several cases from the VQA-Rad dataset due to questions being deemed unanswerable from the provided images. For VQA-Rad we report the accuracy for the balanced test split and the test split suggested by Xu et al. (2023). In pathology, it showed reasonable performance on this useful, albeit noisy dataset, as measured by accuracy for Yes/No questions and average tokenized F1-score for zero-shot responses.

| Dataset | Subset | Metric(%) | Med-Gemini | Gemini Ultra | SoTA |
|------------------------|---|-----------------------------------|--------------|--------------|---|
| VQA-Rad | Test set from Xu et al. (2023) (CXR only) | Expert score (Excluding 12 cases) | 71.9 | - | 57.9 (Xu et al., 2023) |
| | | Accuracy (Closed only) | 78.8 | - | 67.1 (Xu et al., 2023) |
| | Balanced Test (CXR only) | Expert score (Excluding 4 cases) | 71.8 | - | 55.7 [‡] (Xu et al., 2023) |
| | | Tokenized F1 | 62.6 | 49.0 | - |
| | | Accuracy (Closed only) | 78.1 | 62.4 | - |
| | Balanced Test (CXR, CT, MR) | Expert score (Excluding 44 cases) | 61.9 | - | - |
| | | Tokenized F1 | 50.1 | 46.4 | - |
| | | Accuracy (Closed only) | 69.7 | 64.9 | - |
| | Slake-VQA | Official Test (English only) | Tokenized F1 | 75.8 | 44.6 |
| Accuracy (Closed only) | | | 84.8 | 70.4 | 91.1 (Li et al., 2023b) |
| PathVQA | Official Test | Tokenized F1 | 58.7 | 34.9 | 62.7[†] (Tu et al., 2024) |
| | | Accuracy (Closed only) | 83.3 | 62.8 | 90.9 (Sun et al., 2024) |
| MIMIC-CXR VQA | Test split from Xu et al. (2023) | Tokenized F1 | 52.5 | 44.8 | - |
| | | Accuracy (Closed only) | 78.6 | 70.9 | 68.1 (Xu et al., 2023) |

[†] Using one-shot prompt with a text-only exemplar (Tu et al., 2024).

[‡] Average of original scores over new, smaller test set.

4.2. Visual question answering (VQA)

We assessed Med-Gemini-2D’s performance on VQA tasks across a range of diverse medical specialties, including radiology, dermatology, and pathology, and spanning a wide range of open-ended and closed-ended questions. Table 6 summarizes the overall VQA results. Prompt templates were manually optimized for each model and VQA dataset on the validation splits, and are listed in Table A.2. Model answers were generated using the same method and parameters as for CXR classification, see Section 4.1. That is, the generative sampling was not constrained by a given test vocabulary in any manner, as it was in related work (Li et al., 2023a,b; Zhang et al., 2023a), typically only to the test set’s ground truth answers, for the reasons described e.g. in Tu et al. (2024); Van Sonsbeek et al. (2023). In other words, answers were generated in a truly generative, open-ended, and zero-shot manner.

For close-ended questions, we measured accuracy based on exact matches of normalized model vs ground truth answers, and compare against SOTA based on vocabulary-constrained answer generations, since the freely generated answers mostly matched the overall set of ground truth answers. For open-ended questions, we report the average token-wise F1 score (Tu et al., 2024) between the normalized answers of the model and the ground truth, and only compare against SOTA results where answers were generated in the same zero-shot manner. In addition, for Med-Gemini-2D results on VQA-Rad, one board-certified radiologist scored the answers using the 3-point scoring rubric introduced in Xu et al. (2023), in order to compare against results of the prior ELIXR model.

We assessed Med-Gemini-2D’s VQA capabilities in radiology using three datasets from distinct domains. First, we evaluated on the MIMIC-CXR VQA test set, an in-distribution benchmark containing 226 question-answer pairs for 48 chest X-ray images suggested by Xu et al. (2023). Second, we used the English-only 1,061 question and answer pairs in the test split of Slake VQA, a large bilingual (English and Chinese) VQA dataset. Finally, we employed the VQA-Rad dataset, leveraging the new three-way balanced split detailed in Section 2. As discussed previously, to evaluate out-of-distribution performance and facilitate a head-to-head comparison with the previous best-in-class model (ELIXR), we did not fine-tune our model on either the VQA-Rad training images or questions. This approach ensures both ELIXR and our model are tested on the same, larger VQA-Rad test set (Xu et al., 2023). Moreover we evaluated the performance of our model on both chest X-ray only and all modality (CT, MRI, and X-ray).

As Table 6 demonstrates, Med-Gemini-2D outperformed many previous results and Gemini across different subsets and metrics. Specifically, in the chest X-ray only (ELIXR split) subset, our model achieved a remarkable expert-evaluated accuracy score of 71.9 and an accuracy of 78.8 in closed-ended questions, improving the best-in-class number by 14 and 11.7, respectively. In the chest X-ray-only balanced split subset, our model maintained strong performance with an expert-evaluated accuracy of 71.8 improving best-in-class number by 16 and an accuracy of 78.1 in closed-ended questions. Moreover, across all modalities in the balanced split, our model achieved competitive results and improved over Gemini, demonstrating its versatility. In the MIMIC-CXR VQA dataset, it achieved an accuracy of 78.6% on Yes/No questions and a tokenized F1 score of 52.5 overall. In the Slake VQA dataset, our model significantly outperformed Gemini and achieves performance close to state-of-the-art with 84.8 accuracy on close-ended questions, showcasing its capability across different domains. Its mean tokenized F1-score across all English questions at 75.8 is lower than for the MedPaLM-M model (Tu et al., 2024) at 89.3, which might be partially attributable to it being prompted in a zero-shot manner, versus a one-shot text-only prompt for the latter. Contrary to MedPaLM-M, Med-Gemini-2D was not fine-tuned with one-shot examples, hence this prompting technique would yield worse results during inference.

To evaluate Med-Gemini-2D’s VQA capabilities in pathology, we utilized the PathVQA dataset (He et al., 2020). For this dataset, our model achieved an accuracy of 83.3 at Yes/No questions, and a tokenized F1-score of 58.7 over all questions which improves over Gemini. Its overall zero-shot results are slightly below those of the MedPaLM-M model (Tu et al., 2024), which reports an overall tokenized F1-score of 62.7, albeit employing a text-only one-shot prompting technique here as well. This technique involved an additional exemplar question-and-answer pair along with an image placeholder string `` provided as a one-shot example during evaluation. However, while these results can provide a general sense of VQA capabilities for images comprising both histopathology and general anatomic pathology photographs and diagrams, given the known issues with QA pairs and image quality in this auto-generated dataset (Lu et al., 2024), we suggest cautious interpretation.

We also conducted a qualitative review of model behavior for histopathology and radiology VQA tasks. Examples are shown in Figure 6, and Figure 7 .

Table 7 | Evaluation details report generation in chest X-rays. Med-Gemini-2D sets a new standard for AI generated chest X-ray (CXR) report generation based on expert evaluation, exceeding previous best results across two separate datasets.

| Dataset | Metric | Med-Gemini | SoTA |
|-----------|--|------------|--------------------------|
| MIMIC-CXR | AI superior or similar to original report (Normal) | 57% | 45% (Tanno et al., 2024) |
| | AI superior or similar to original report (Abnormal) | 43% | 42% (Tanno et al., 2024) |
| | Clinically acceptable AI generated report (All) | 72% | - |
| IND1 | AI superior or similar to original report (Normal) | 96% | 85% (Tanno et al., 2024) |
| | AI superior or similar to original report (Abnormal) | 65% | 53% (Tanno et al., 2024) |
| | Clinically acceptable AI generated report (All) | 88% | - |

4.3. Report generation for chest X-rays

In clinical practice, the role of the radiologist extends far beyond narrow interpretation of radiology images. Radiologists are tasked with conveying nuanced findings within a broader clinical context, synthesizing information, and providing recommendations for patient care. Expert radiologists use natural language to articulate this synthesis of imaging findings, overall impressions, and recommendations in written reports. Unlike some prior work, our model was tuned for the difficult task of generating both the ‘FINDINGS’ and ‘IMPRESSION’ sections of chest X-ray reports for frontal view chest radiographs (anterior-posterior or posterior-anterior), covering comprehensively both the observations and inferences typically made by radiologists during a study.

Table 8 presents the performance comparison of various models in generating radiology reports for chest X-rays using the publicly available MIMIC-CXR dataset. The “Sections” column indicates whether the model generates the ‘FINDINGS’ (‘F’) or ‘IMPRESSIONS’ (‘I’) section of the report, with metrics drawn from published research. Higher values in all metrics indicate superior performance. Notably, our model undertakes the more challenging task of generating both sections (F + I) for frontal chest X-rays, aiming to capture the radiologist’s holistic interpretation of the study.

Following common practice, we leveraged the established n-gram based methods such as ROUGE-L, BLEU-4 to evaluate the generated reports quality against the ground-truth. Additionally, we measured the RadGraph F1-score, which is the F1 score between the entities extracted from the reference report and generated one using RadGraph (Jain et al., 2021). RadGraph accounts for not only the absence or presence of findings in the report, but also their relationships to image features. Med-Gemini achieved a RadGraph F1-score of 24.4%, marking a notable improvement of 3.9% compared to the previous top-performing model.

For the IND1 dataset we did not evaluate automated metrics, as automated metrics such as RadGraph F1-score are specifically trained on MIMIC-CXR to measure performance of US-style chest X-ray report and are not capable of handling the out-of-distribution format of IND-1 dataset reports obtained in an India-based clinical setting.

Human evaluation rubric for report generation For report generation we devised a novel evaluation rubric, expanding on those used in Flamingo-CXR (Tanno et al., 2024) and Med-PaLM M (Tu et al., 2024), to understand potential impact on clinical management of patients. The evaluation rubric consists of six categories that compare two reports. It provides an improved granularity around patient impact and was used for both the CXR and CT generated reports. Table 9 defines the labels for comparing the AI and original radiologist reports for the same study. This rubric along with training materials and examples were provided to radiologist labelers as training material, prior to any labeling. In each example seen by labelers, the origin of the reports (AI vs. original) was masked and the reports were shown in random order to avoid bias.

Table 8 | Automated report generation metrics on the MIMIC-CXR dataset. This table presents the performance of various models on generating radiology reports for chest X-rays using the publicly available MIMIC-CXR dataset. The *Sections* column indicates whether the model generates the FINDINGS (F) or IMPRESSION (I) section of the report, with metrics sourced from published research. For all of the metrics higher is better. Bold values highlight the best results in each category for (F + I) methods. Notably, our model tackles the more challenging task of creating both sections (F + I) for frontal chest X-rays (anterior-posterior or posterior-anterior views), aiming to capture the radiologist’s comprehensive interpretation of the study. Med-Gemini achieved a RadGraph F1-score of 24.36% on chest X-ray report generation, demonstrating a 4.0%+ improvement over the previous best-in-class score.

| Model | Sections | NLG Metrics(%) | | | Clinical Metrics(%) |
|---|----------|----------------|-------------|-------------|---------------------|
| | | CIDEr | BLEU4 | Rouge-L | RadGraph F1-score |
| CXR-RePaiR (Endo et al., 2021) | F | - | 2.1 | 14.3 | 9.1 |
| M^2 Transformer (Miura et al., 2020) | F | 50.9 | 11.4 | - | 22.0 |
| RGRG (Tanida et al., 2023) | F | 49.5 | 12.6 | 26.4 | - |
| METransformer (Wang et al., 2023a) | F | 36.2 | 12.4 | 29.1 | - |
| Med-PaLM M, 12B (Tu et al., 2024) | F | 23.4 | 10.4 | 26.2 | 25.2 |
| Med-PaLM M, 84B (Tu et al., 2024) | F | 26.2 | 11.3 | 27.3 | 26.7 |
| MAIRA-1 (Hyland et al., 2023) | F | - | 14.2 | 28.9 | 24.3 |
| R2Gen (Chen et al., 2020) | F + I | - | 10.3 | 27.7 | 13.4 |
| WCT (Yan et al., 2021) | F + I | - | 14.4 | 0.274 | 14.3 |
| CvT-21DistillGPT2 (Nicolson et al., 2023) | F + I | 36.1 | 12.4 | 28.5 | 15.4 |
| BioVil-T (Bannur et al., 2023) | F + I | - | 9.2 | 29.6 | - |
| R2GenGPT (Wang et al., 2023b) | F + I | 26.9 | 13.4 | 29.7 | - |
| Flamingo-CXR (Tanno et al., 2024) | F + I | 13.8 | 10.1 | 29.7 | 20.5 |
| Med-Gemini-2D | F + I | 17.5 | 20.5 | 28.3 | 24.4 |

Five India-based board-certified radiologists, one India-based thoracic specialist, and one US-based academic thoracic radiologist evaluated a total of 606 cases: 306 from the MIMIC test dataset and 300 from IND1. After the study completion, readers were compared using their mean Quadratic Kappa (Sim and Wright, 2005) to the two thoracic specialists. Two readers falling below 0.2, i.e. “none to slight agreement” were eliminated from the final results. The results were computed based on the total sum of categories for the selected reports after elimination of scores of the X category. The percentage of cases within each category were then plotted sequentially along a horizontal plot for all, abnormal, and normal cases as shown in Figure 3 and summaries are shown in Table 7

In examining cases that fell into the **A1** and **B1** categories, i.e. where one report captures clinical findings but both would result in the same patient management, similar reasoning was given in both categories. These included missing less critical findings and descriptiveness of findings. Examples of missed findings include: mild cardiomegaly, calcified granulomas, and old fractures. In terms of descriptiveness, examples include: better descriptions of bulla, proper identification of devices, and clearly discerning mass versus pneumonia and other less explicit diagnoses. Reports falling into categories **A2** and **B2** missed key findings, including: failures in assessing tube positions, missed nodules, and missed pneumothraces.

4.4. Report generation for head/neck CT volumes

3D imaging modalities often involve more complex data preparation and longer radiologist interpretation time in comparison to 2D images such as X-rays, making the paired image-text data required for generative AI modeling scarcer and more expensive. Additionally, radiology reports tend to be much longer and imaging features much sparser for 3D images than for 2D images. Given this relative

Table 9 | Human evaluation rubric comparing AI generated radiology reports to original reports.

| Rubric Score | Rubric Definition |
|--------------|---|
| A2 | Report A captures key clinically relevant findings that are not found in B. Report A would result in correct patient management and report B would not. |
| A1 | Report A captures more relevant findings, but both would result in the same correct patient management. |
| C | Both reports capture similar findings in the image and would result in correct patient management. |
| B1 | Report B captures more relevant findings, but both would result in the same correct patient management. |
| B2 | Report B captures key clinically relevant findings that are not found in A. Report B would result in correct patient management and report A would not. |
| X | Neither report would result in correct patient management. |

data scarcity and information complexity (with correspondingly increased memory requirements), end-to-end modeling to convert 3D radiology images to text reports has previously been infeasible. With its increased computational capacity and extensive domain-specific pretraining, Med-Gemini-3D, building on other recent generative AI work such as [Hamamci et al. \(2024\)](#), is the first LLM-based generative AI model able to interpret a 3D medical imaging modality end to end from the CT volume to text.

Using the same human evaluation rubrics introduced in Section 4.3, we evaluated a total of 92 non-contrast head/neck CT studies consisting of 27 Normal-labeled cases without findings and 65 Abnormal-labeled cases that contained findings, including both acute findings such as cerebrovascular accidents as well as findings that are common consequences of aging, such as atrophy. Studies were initially divided into normal and abnormal candidates based on the length of the impression sections of the reports. A random subset within each was selected and then manually classified by a board-certified radiologist into the normal or abnormal category based on the full radiology report. Studies classified as normal contained no findings.

In reviewing the reports, a single academic board-certified examined the study and all series using a web-based Picture Archiving and Communication System (PACS) viewer. The radiologist graded the two reports using the same rubric presented for evaluating CXR reports. For each rating, the radiologist also recorded a comment describing why the rating was given. The model generated the report based on a single series with the most slices and did not have access to any of the other series. The model was given the patient history in the form of text during inference.

Results are shown in Figure 4 and Table 10. We found that 45% of AI reports on normal studies and 57% of AI reports on abnormal studies would have resulted in the correct clinical management of the patient, though some of those AI reports included errors that would not directly affect management. We did find, however, that only 17% of AI reports were considered to be of equivalent or better quality than the original radiologist reports. In examining the notes on errors from the generated reports, *i.e.*, those that were scored **B2**, roughly half involved missed findings while the other half involved hallucinations such as identified subdural hematomas or cysts. In terms of **B1** category reports, comments about the generated report mention it either incorrectly estimates or under-characterizes white matter changes.

While our early results presented here leave significant room for future improvement, the potential opportunity for AI in volumetric imaging is vast. This difficulty in reporting on volumetric data can result in concerning diagnostic delays ([NHS, 2024](#)). The ability to safely triage, expedite, and quality check existing reports could be highly beneficial in health systems around the world.

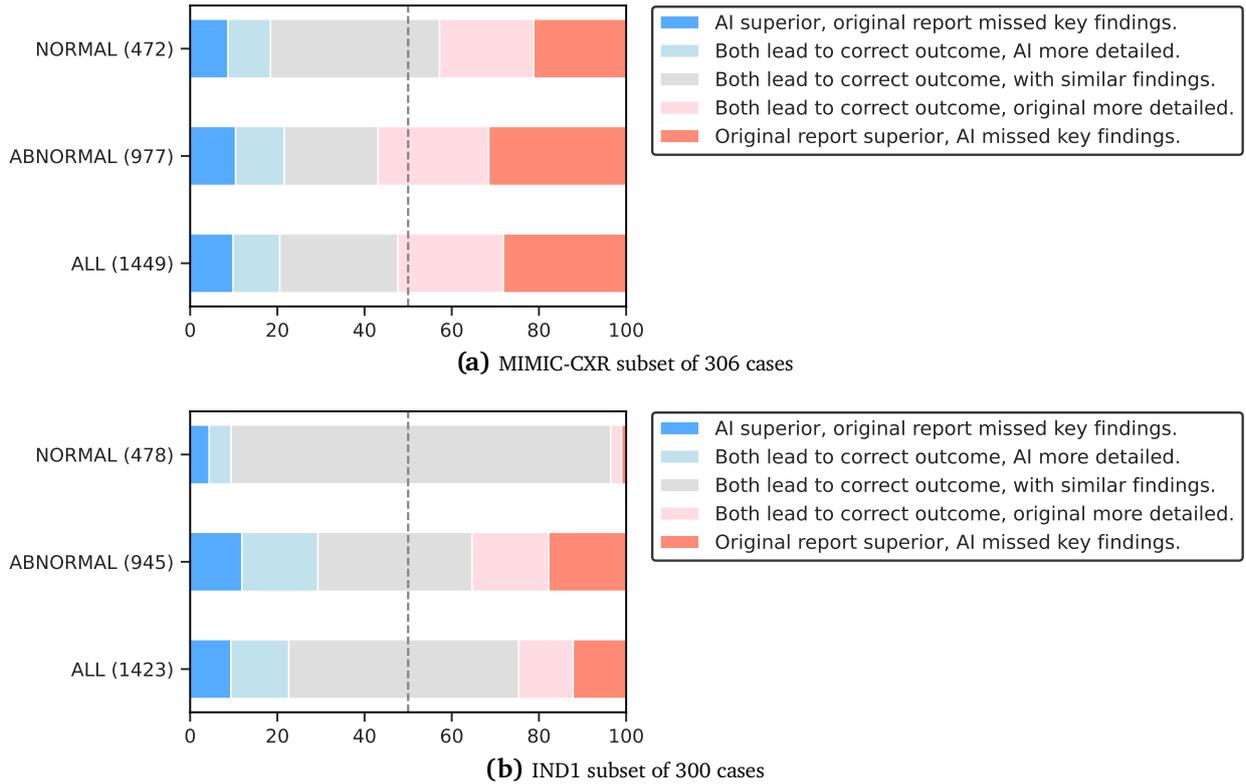


Figure 3 | Med-Gemini-2D CXR report generation results based on 4 India-based board-certified radiologists and one US-based academic board-certified radiologist performing report comparisons. **(a)** a subset of 306 MIMIC-CXR cases. Overall 48% of the cases were equal or superior to the original reports and 72% lead to the same clinical outcome. **(b)** a subset of 300 cases from IND1. Overall 75% of the reports were equal or superior to the original reports. In both studies, the AI performance was better on the normal cases.

Table 10 | **Human evaluation results for Head CT Volume report generation** Note there is no existing best-in-class performance for this task as report generation from Head/Neck 3D CT volumes is a new capability. Additionally, the model had access to a single series in the study for report generation.

| Dataset | Metric | Med-Gemini |
|---------|--|------------|
| CT-US1 | AI superior or similar to original report (Normal) | 15% |
| | AI superior or similar to original report (Abnormal) | 18% |
| | Clinically acceptable AI generated report (All) | 53% |

4.5. Disease prediction from genetic information

Personalized medicine can benefit greatly from genetics, as disease risks depend heavily on an individual’s genetic makeup. To leverage this powerful information, we expanded our model’s ability to process genetic information in the form of an RGB image by featurizing the genome into polygenic risk scores (PRS) as explained in Section 2.

To assess the disease risk prediction capability of Med-Gemini-Polygenic, we created benchmarks by training linear models on all PRS featurizations plus demographics (“Ensemble of PRSs and demographics”) which is the current best practice for using PRSs for disease prediction (Albiñana et al., 2023; Truong et al., 2024). For in-distribution health outcomes (see Section 2) used in Med-Gemini-Polygenic training, we directly applied the trained “Ensemble of PRSs and demographics”

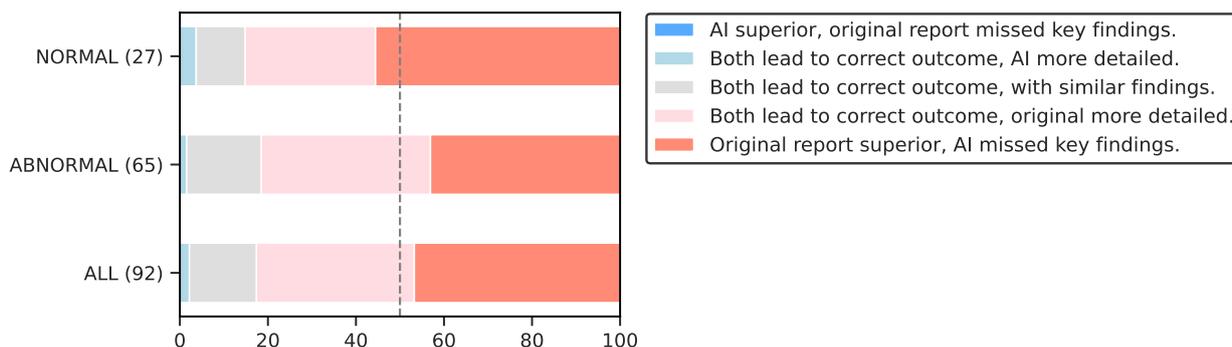


Figure 4 | Med-Gemini-3D Head CT report generation scores on a subset of 92 cases from the CT-US1 test set scored by a US-based board certified radiologist. Across all cases, 17% of the AI generated reports were graded equivalent or superior to that of radiologists’ reports, while 53% were judged as resulting in equivalent patient care. Performance was better overall on abnormal versus normal cases.

models as benchmarks. For health outcomes that were never used in the training process (out-of-distribution or OOD, see Section 2) but share some genetic correlation with the in-distribution outcomes, we first calculated their phenotypic correlations with all the in-distribution outcomes, and used the model trained to predict the most correlated in-distribution outcome to generate a maximally strong performance benchmark.

We evaluated Med-Gemini-Polygenic performance on case/control balanced datasets sampled from the test split (200 cases, 200 controls per outcome) for computational efficiency (Section A.2.2). We obtained a disease probability score from Med-Gemini-Polygenic by prompting it to predict the status of a given health outcome using a text prompt and the genetic risk “image” (Table A.9), and computed the probability as the ratio of the likelihoods of the model generating a positive and negative prediction. Med-Gemini-Polygenic achieved higher AUCs than the PRS linear model benchmarks for all in-distribution health outcomes except glaucoma (Figure 5). To evaluate zero-shot generalization ability, we prompted Med-Gemini-Polygenic to predict disease status for the six out-of-distribution health outcomes. Med-Gemini-Polygenic achieved similar performance to benchmarks trained on the most correlated in-distribution outcome (Table A.11) despite never being instructed about the associations between in-distribution and out-of-distribution outcomes (Figure 5).

Additionally, we compared the performance of linear probes of the Med-Gemini-Polygenic embeddings and directly prompting Med-Gemini-Polygenic in the evaluation sets of 400 individuals. Comparisons of the AUCs show that while Med-Gemini-Polygenic performs similarly to the linear probe when each is only given demographic information, it often outperforms the linear probe when incorporating both PRSs and demographics (Figure A.5). This performance increase is largely attributable to Med-Gemini-Polygenic modeling non-linear interactions between genomic information and demographics (Table A.12).

We caution that the AUC values reported here represent an upper bound on model performance since the GWASs used to create the PRS features were performed within the UK Biobank. However, the relative performance of different models that all operate on this in-sample data is the measure of interest for these analyses.

5. Qualitative Results

In this section we provide a few examples showcasing our model’s capability in medical dialogue for diverse set of medical modalities including chest X-ray, CT, fundus, dermatology, pathology, depicted in Figures 6, 7 as well as 2D (Figure 8) and 3D (Figure 9) radiology report generation. As highlighted

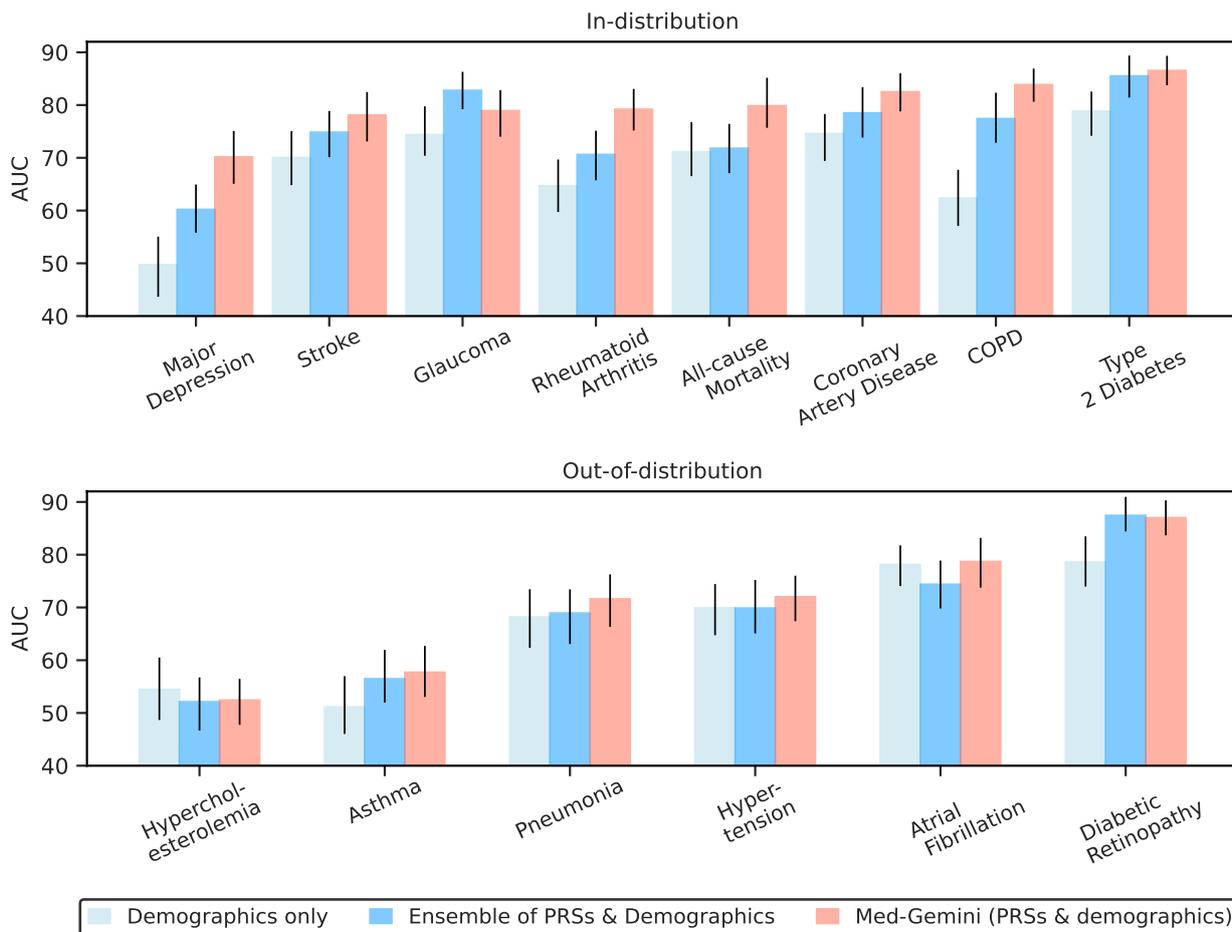


Figure 5 | Health outcome prediction using Med-Gemini-Polygenic compared to two baselines for both in-distribution and out-of-distribution outcomes. “Demographics only” used a linear probe of age, sex, and BMI to predict each health outcome, and “Ensemble of PRSs and demographics” combined demographics with all 7,145 PRSs in a linear probe. Med-Gemini-Polygenic was prompted with both an individual’s PRS image and demographics. For out-of-distribution health outcomes, the linear probes (“Ensemble of PRSs and demographics” and “Demographics only”) were trained to predict the most-correlated in-distribution outcome (Table A.11), and those predictions were then evaluated on the out-of-distribution outcome.

in these examples, Med-Gemini is able to provide accurate and reasonable multimodal dialogue and interpretation capabilities across a variety of medical imaging domains. At the same time, expert review of these examples highlights areas for improvement regarding the phrasing, accuracy, appropriate level of detail, and completeness of generated responses.

In addition to understanding automated report generation capabilities, it is important to consider plausible real world assistive use cases. As a proof of concept, we experimented with directing the model’s attention to a specific region/organ within the CXR (Figure 10).

Lastly, as shown in the above examples, even though Med-Gemini was only fine-tuned with data directly related to image interpretation (e.g. there were no question-answer pairs related to treatments or symptoms in the fine-tuning set), Med-Gemini can still leverage the medical knowledge from Gemini pretraining to give simple but reasonable answers to those questions. While we emphasize that real-world medical diagnosis, prognosis, and treatment information is much more complicated and nuanced than the examples provided here, these examples serve as a proof of concept for combining large model pretraining with domain specialization, an active area for further improvements.

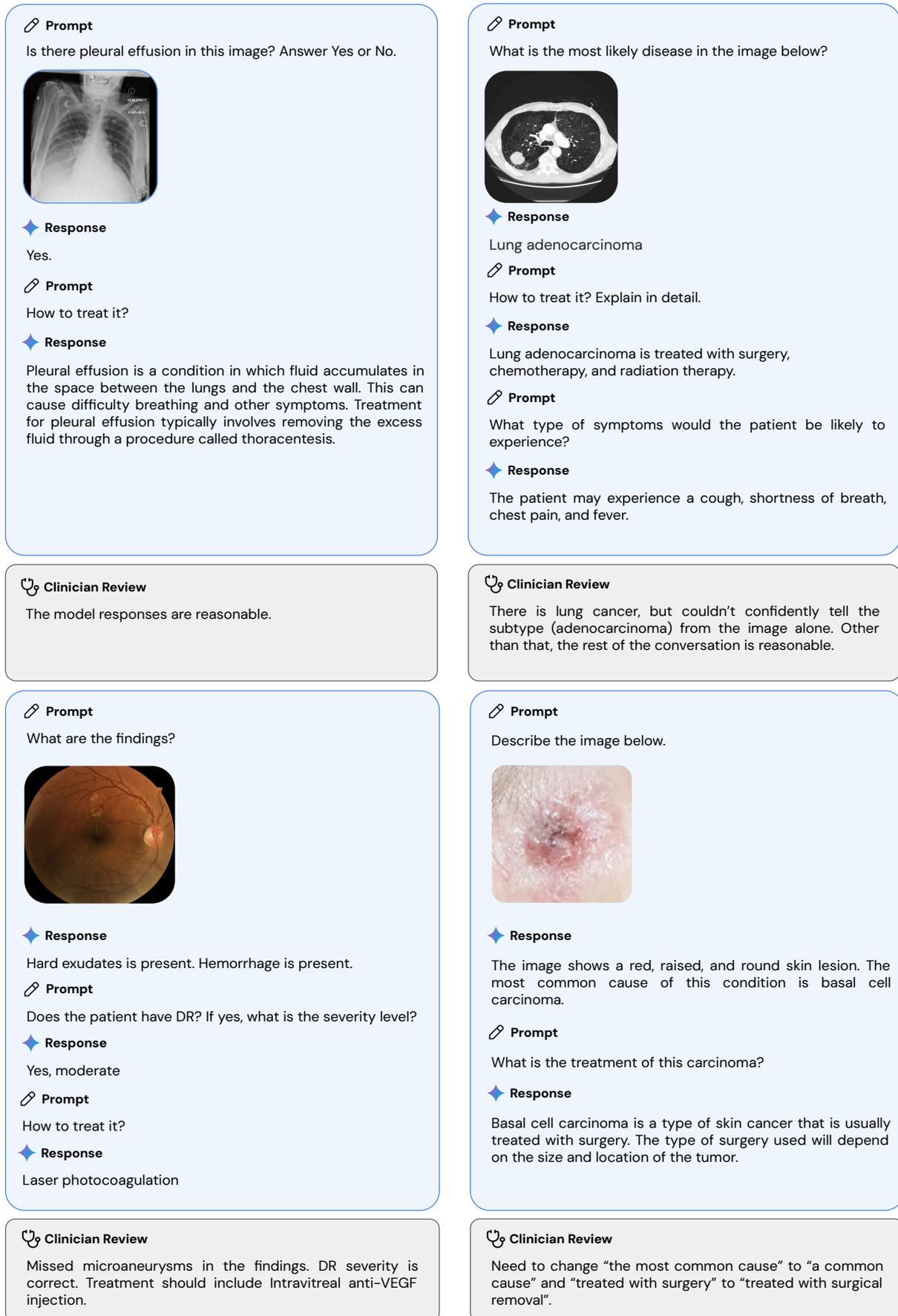


Figure 6 | Example of 2D medical image dialogue via open ended question answering. For chest X-ray (Johnson et al., 2019a), lung CT (Liu et al., 2021), fundus images (Cuadros and Bresnick, 2009), and skin lesion images (Pacheco et al., 2020).

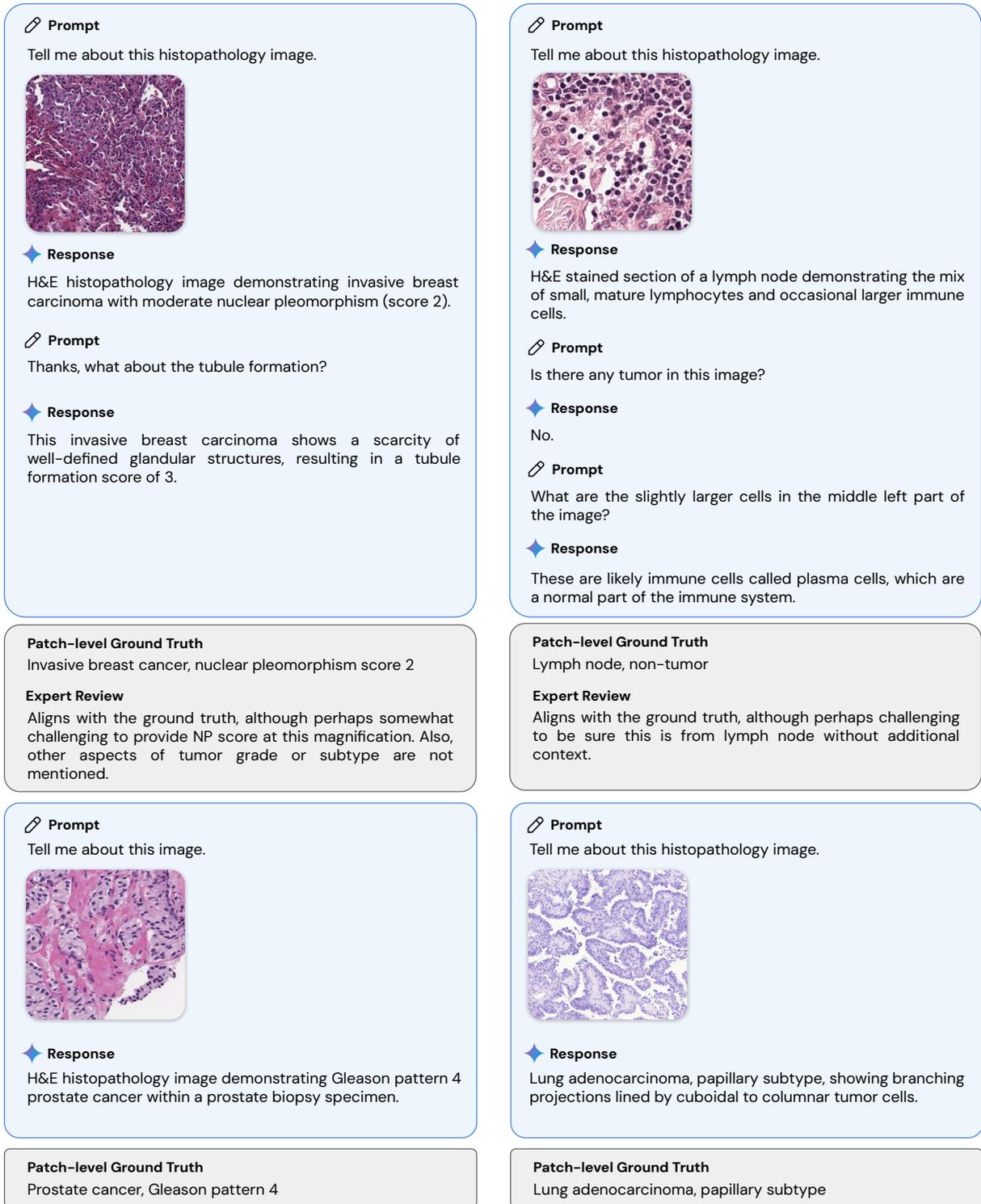


Figure 7 | Examples of histopathology image-based dialogue. These examples highlight accurate histopathology interpretation and communication of information across a range of tissues and findings with only a small amount of visual context. While there is some promising initial evidence for ability to interact and reason further about images (top right), current capabilities for follow up interaction and exploration of input images are limited and this remains an active area for improvement. This demonstration utilizes histopathology patches from TCGA, CAMELYON16, and other data sources as described in [Lai et al. \(2023\)](#).

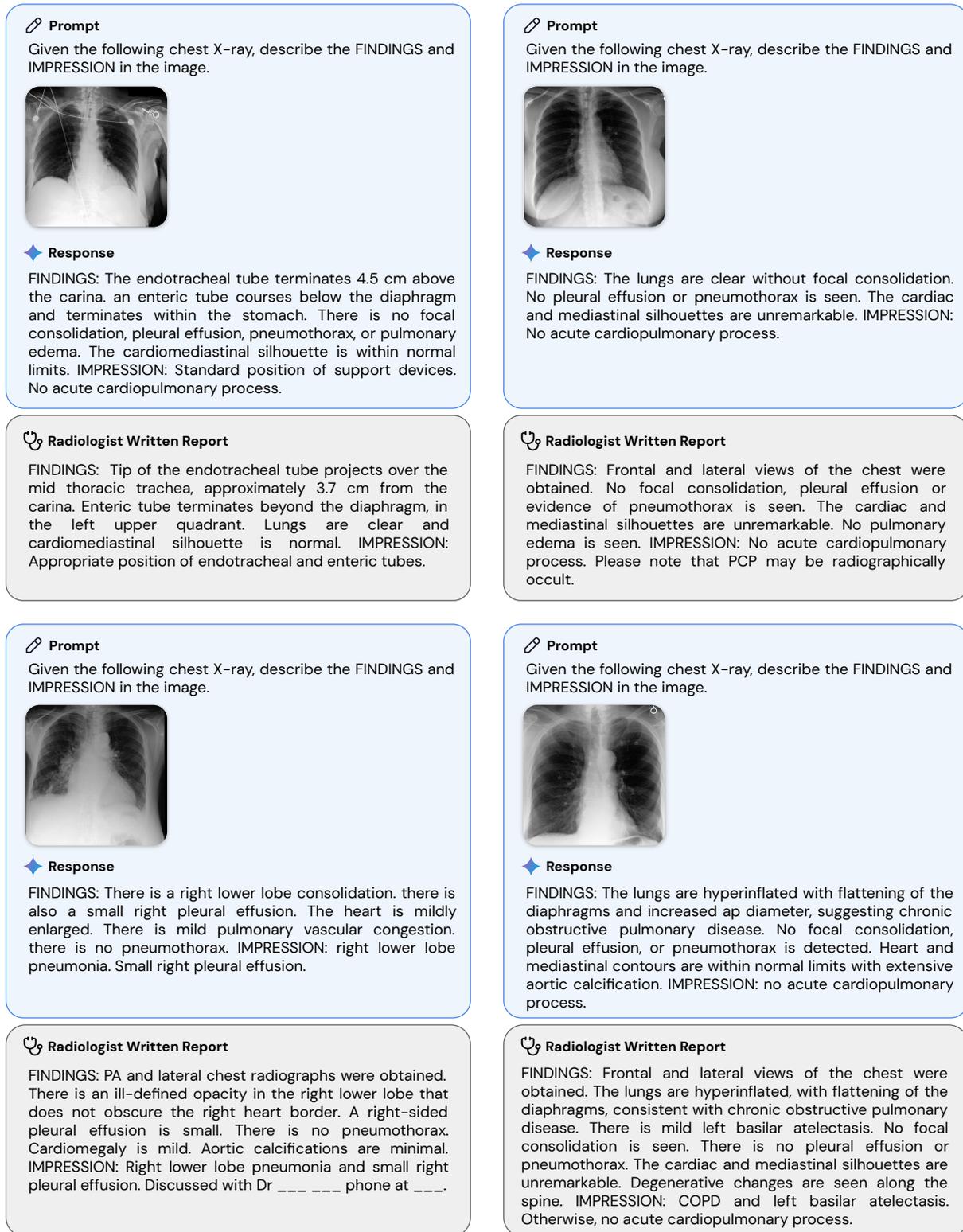
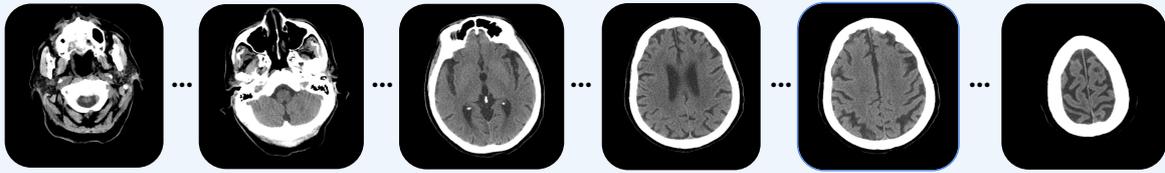


Figure 8 | Examples of chest X-ray report generation. These examples demonstrate the capability of Med-Gemini for CXR report generation on various conditions. Top-left, support devices; Top-right, normal case; Bottom-left, acute abnormality; Bottom-right, chronic abnormality. A radiologist reviewed all these examples and confirmed model generated reports are reasonable with one note for the bottom-right case where the “increased ap diameter” is usually detected from a lateral image, and thus a hallucination from the model.

 Prompt

Given the image, what are the radiology findings?



 Response

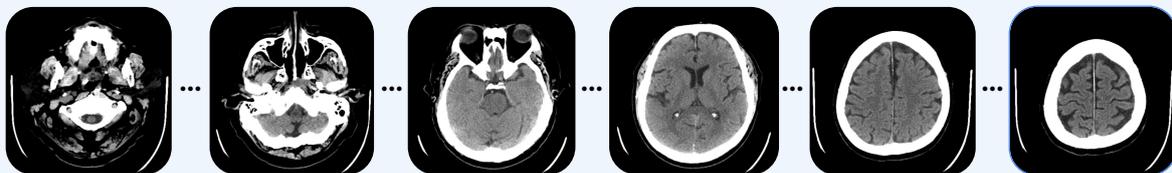
FINDINGS: mild cortical atrophy and periventricular white matter changes are noted compatible with microvasculopathy. there is proportionate ventriculomegaly. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect. no abnormal extra-axial fluid collection is identified. no definitive evidence of intracranial hemorrhage or large territorial infarct within the limits of ct imaging. old left basilar ganglia infarct. no intracranial mass, swelling, hemorrhage, infarction or extra-axial collection is seen. visualized mastoid air cells: well pneumatized. visualized paranasal sinuses: well pneumatized. atherosclerotic calcification of vasculature at the skull base.

 Radiologist Written Report

FINDINGS: the fourth ventricle is midline. the remaining ventricular system is within normal limits. there is mild chronic white matter microangiopathic disease and moderate generalized cortical atrophy. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. bone windows demonstrate an intact skull. the remaining visualized paranasal sinuses and mastoids are clear. chronic white matter microangiopathic disease, mild generalized cortical atrophy, not significantly changed since 2015. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect.

 Prompt

Given the image, what are the radiology findings?



 Response

FINDINGS: there is a moderate amount of scattered low density in the periventricular and subcortical white matter of the cerebral hemispheres. there is no evidence of intracranial mass effect, acute hemorrhage, midline shift or extra-axial fluid collection. no acute subarachnoid, subdural, epidural, or intraparenchymal hemorrhage is visualized. there is no abnormal intracranial mass effect, midline shift, hydrocephalus, acute hemorrhage, or large infarct. no hydrocephalus, subarachnoid hemorrhage, or mass. no mass, hemorrhage, or acute infarction is demonstrated. cerebellum: no edema, hemorrhage, mass, acute infarction, or inappropriate atrophy. brainstem: no edema, hemorrhage, mass, acute infarction, or inappropriate atrophy. sella: no parasellar mass identified. skull: no mass or significant visible lesion. the calvarium is intact. the mastoid air cells and middle ear cavities are normally aerated. there is minimal mucosal thickening lining the right and left maxillary sinuses. minimal mucosal thickening lining the right and left maxillary sinuses.

 Radiologist Written Report

FINDINGS: scattered punctate periventricular and subcortical white matter hypodensities without mass effect or volume loss are compatible with mild microangiopathic white matter changes. ventricles and sulci are normal in size and configuration. no parenchymal hemorrhage, intra-axial or extra-axial fluid collection, or mass lesion is present. no acute transcortical infarction, regional mass effect, transtentorial herniation, or midline shift is present. visualized paranasal sinuses are clear. visualized mastoid air cells are clear. visualized osseous labyrinth structures appear normal. visualized orbits are normal. visualized soft tissues of the scalp are normal. no calvarial fracture. skull base and craniocervical junction are normal. atherosclerotic calcification tracks along the cavernous and supraclinoid internal carotid artery segments.

Figure 9 | Examples of 3D Head CT report generations. These examples showcase 3D medical image dialogue for Head CT report generation: (top) correct abnormal case, (bottom) incorrect abnormal case. While Med-Gemini can identify some abnormalities missed by radiologist generated reports (highlighted in green), it can also mischaracterize findings that are present (highlighted in red) or hallucinate findings that are absent from the image.

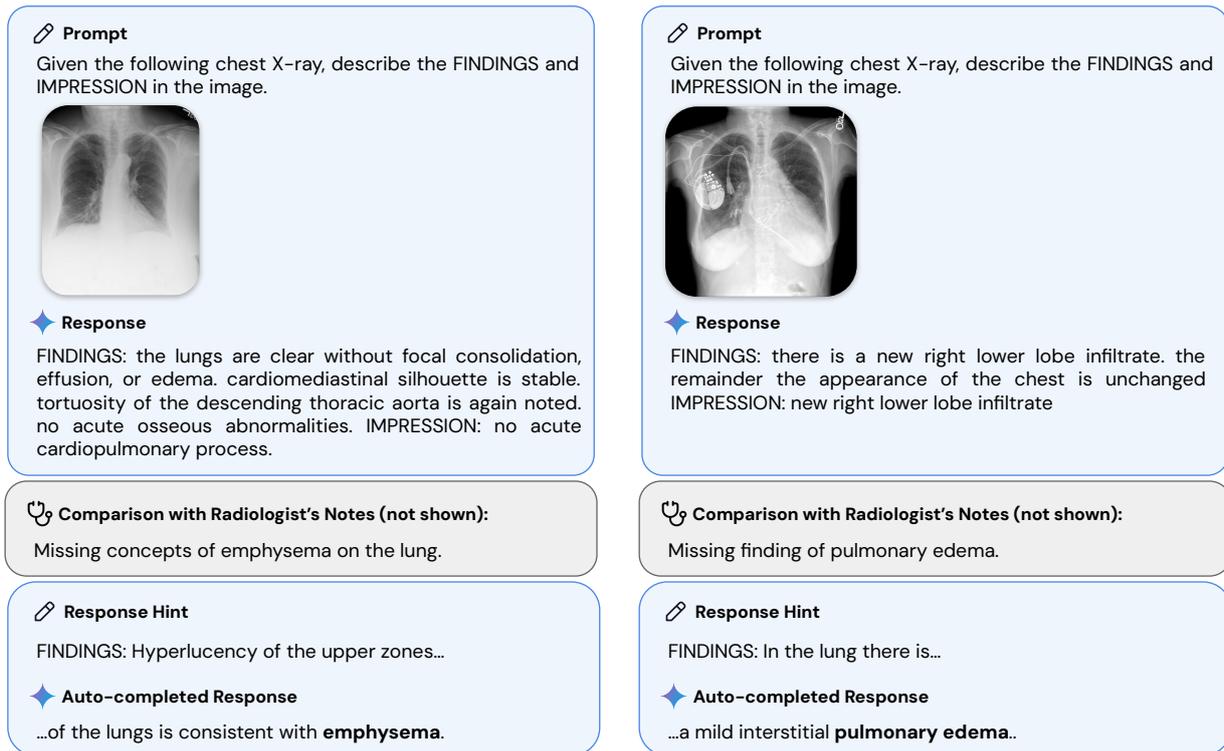


Figure 10 | Examples of chest X-ray report auto-completion. In these examples, particular concepts were missing from the report generated without any hint, and were recovered with the autocomplete prefix hint. A) Emphysema, B) Pulmonary Edema.

6. Related Work

The evolution of medical language models Large language models (LLMs) built on Transformer architectures (Parmar et al., 2018; Vaswani et al., 2017) have seen rapid advancement, driving significant progress in natural language processing and multimodal modeling. Pathway scaling methods (Barham et al., 2022) have been crucial in enabling the development of ever-larger models like the PaLM family including PaLM, PaLM 2, and PaLM-E (Anil et al., 2023; Chowdhery et al., 2023; Driess et al., 2023). Other significant LLMs include BERT (Devlin et al., 2018), GPT family (Achiam et al., 2023; Brown et al., 2020; Radford et al., 2019), T5 (Raffel et al., 2020), and LLaMA (Touvron et al., 2023), Hyena (Poli et al., 2023), Mistral 7B (Jiang et al., 2023). LLMs are often refined through techniques like Chain of Thought (CoT) prompting (Wei et al., 2022) or fine-tuning (FLAN) (Wei et al., 2022).

These advancements have catalyzed an expansion of LLMs specifically designed for medical domains, such as PubMedGPT (Bolton et al., 2022), BioGPT (Luo et al., 2022), Med-PaLM (Singhal et al., 2023a) and its successor Med-PaLM 2 (Singhal et al., 2023b), Clinical Camel (Toma et al., 2023), MedAlpaca (Han et al., 2023), BioMistral (Labrak et al., 2024), LLMs for clinical trial recruitment (Wornow et al., 2024), and others. Language models can handle omic information, as demonstrated by models such as HyenaDNA (Nguyen et al., 2024), BioT5 (Pei et al., 2023), sc-GPT (Cui et al., 2024), and ProtLLM (Zhuo et al., 2024).

Multimodal models in medicine Beyond language and text alone, multimodal models like Flamingo (Alayrac et al., 2022), PaLI (Chen et al., 2022), GPT-4 (Achiam et al., 2023), GPT-4v (OpenAI, 2023), and LLaVa (Liu et al., 2023, 2024a) have demonstrated remarkable capability in processing both text and images. Gemini (Gemini Team, Google, 2023; Google, 2024) introduced further advancement in

multimodal capabilities, exhibiting a distinct ability to reason across text, images, and other modalities such as video and audio.

Building upon these capable generic multimodal models, for medical applications specifically, recent works include vision-language models that span multiple medical imaging modalities as well as those that focus on a specific imaging domain, such as radiology or histopathology. Efforts such as Med-Flamingo (Moor et al., 2023b), BiomedCLIP (Zhang et al., 2023a), Med-PaLM M (Tu et al., 2024), BiomedGPT (Zhang et al., 2023a), Flamingo-CXR (Tanno et al., 2024), LLaVa-Med (Li et al., 2024), PMC-VQA (Zhang et al., 2023b), RadFM (Wu et al., 2023), ELIXR (Xu et al., 2023), XrayGPT (Thawkar et al., 2023), MAIRA-1 (Hyland et al., 2023), HeLM (Belyaeva et al., 2023), M-REGLE (Zhou et al., 2024), CONCH (Lu et al., 2024), PLIP (Huang et al., 2023), PathAsst (Sun et al., 2024), QuiltNet-B-32 (Ikezogwo et al., 2024) and many others specifically explore the potential of multimodal models for medical applications, signaling a growing interest in this area. These methods cover a range from generalist to specialist approaches. Models such as MAIRA-1 (Hyland et al., 2023), XrayGPT (Thawkar et al., 2023), Radiology-GPT (Liu et al., 2024b), and CT2Rep (Hamamci et al., 2024) focus on radiology report generation, and among modalities choose only chest X-ray or chest CT report generation. Some of these approaches broaden their capabilities to cover multiple types of modalities but focus on only one task, such as methods that aim for VQA capabilities like LLaVA-Med (Li et al., 2024), and PMC-VQA (Zhang et al., 2023b), aiming to build assistants for medical question answering.

While specialized VLMs demonstrate particular strengths, generalist models capable of handling a wide range of tasks and modalities, such as Med-PaLM M, are gaining prominence. The field of medical AI is witnessing the emergence of comprehensive ‘Generalist Medical AI’ models (Moor et al., 2023a,b; Tu et al., 2024; Zhang et al., 2023a) and the orchestration of AI tools for medical tasks using LLMs (Ferber et al., 2024). These models aspire to provide robust interaction with medical information in a manner similar to what general-purpose LLMs have done for broader domains. Pioneering efforts like these offer important initial insights into the potential for large multimodal models to provide assistance across various medical tasks using a unified platform. This inconsistency underscores the urgent need for a unified benchmark to enable meaningful evaluation in this rapidly evolving field.

Multimodal evaluation benchmark and metrics Evaluation of medical VLMs suffers from a lack of consistency and standardization, creating a new landscape for works proposing new benchmarks to fill this gap. Multiple recent works demonstrate this inconsistency with varying tasks, datasets, and completely distinct sets of metrics, hindering direct comparison even for a same dataset. Along these lines, multiple recent works (Fleming et al., 2023; Moor et al., 2023b; Royer et al., 2024; Tu et al., 2024; Wu et al., 2023) suggest multimodal benchmarks such MultiMedEval, MultiMedBench, RadBench, RadMD, and MedMD to evaluate these generalist and multimodal models in a more systematic fashion.

7. Discussion

In this study, we present three new models within the Med-Gemini family, based upon Gemini 1.5, across various medical modalities. We show promising performance across a number of tasks, including classification, VQA, and report generation. Our Med-Gemini models are able to process complex medical data types, including 2D and 3D radiology images, histopathology patches, ophthalmology images, dermatology images, and genetic risk scores. Importantly, our models were fine-tuned using predominantly medical data and paired free text descriptive reports. These reports are ubiquitous in healthcare and our ability to use them as a training objective reduces the need for further expensive expert labelling.

The results in this study show early potential across a number of different tasks and individual

modalities. We believe that the combination of tasks and multiple modalities in future work will enable AI models to address a far wider range of applications than has been previously possible. Longer context windows and improved reasoning abilities will enable decision-making that incorporates historical context, more closely reflecting how human specialists operate.

The opportunity for LMMs to analyze complex medical types including 3D radiology and large pathology images presents an exciting range of potential downstream applications. This work showcases our early explorations in CT, a three-dimensional modality that has been challenging to integrate with LMMs to date. This is due to a combination of vast data size, architectural limitations, and the jump in clinical task complexity of interpreting 3D imaging modalities (vs. 2D). While our results are currently a proof of concept, and do not yet reach performance required for clinical use, we expect architectures to rapidly improve. We look forward to exploring other similar complex modalities in future work.

While our findings in this study are promising and provide a glimpse into the potential of LMMs in medicine, it is important to thoroughly test them beyond traditional academic benchmarks. This is necessary to ensure they are safe and reliable before considering use in real-world situations, especially in safety critical areas like healthcare. In this work, we have tried to go deeper into the nuance of medical evaluation through the use of panels of specialists to assess and rate the performance of models on tasks such as report generation and question answering. We believe that an increasingly diverse range of healthcare professionals need to be deeply involved in future iterations of this technology, helping to guide the models towards capabilities that have valuable real world utility. There are a number of areas on which future evaluations should focus before models like these are considered safe and effective for clinical use:

Closing the gap between benchmark and bedside Despite the potential of machine learning in healthcare, there is growing concern about the reliability of algorithm validation methods. In medical image analysis, improvement on simple benchmark performance metrics may not translate to improved outcomes in clinical settings, leading to a disconnect between expectations and real-world usefulness. Benchmark datasets are an important step towards developing clinically useful models, but given their limitations in size, scope, and reflection of real world distributions, they are not themselves a proxy for real-world performance. The potential for generative AI lies foremost in assisting, rather than replacing human specialists in the diagnosis and management of disease; evaluations should shift from static benchmarks to realistic clinical scenarios that assess AI-human collaboration and its impact on patient outcomes.

Identifying and mitigating data bias and safety risks LLMs and LMMs trained on vast datasets risk inheriting biases and errors from their source data. This can lead to misdiagnoses and amplification of systemic bias. Before models like these are used in real world settings, careful evaluations that address safety and bias risks should be performed and any discovered risks should be mitigated (Weng et al., 2024). End users should also carefully validate model performance for their specific use cases and patient populations.

Minimizing data contamination when evaluating zero-shot generalization in large models While LLMs exhibit impressive zero-shot generalization, it's important to note that their massive training datasets increase the potential for data contamination, which may result in overestimation of their true generalization abilities. Large models like Gemini might have inadvertently “seen” examples related to the task during training, even if those examples were not explicitly labeled. This hidden exposure compromises our understanding of models’ true ability to generalize to completely novel concepts when evaluating on open datasets. Researchers are actively investigating the impact of data contamination to ensure we accurately gauge capabilities of such large models (Udandarao et al., 2024; Vogel et al., 2022). Prospective studies, while typically more expensive and time-consuming to execute than retrospective studies, are another option for mitigating this risk.

8. Conclusion

Multimodal generative AI, exemplified by powerful models like Gemini, holds great potential to revolutionize healthcare. While medicine is a rapidly growing use case for these new models, general purpose models may not naturally perform well in the medical domain due to its highly specialized data.

To explore the potential for models like Gemini in medicine, we developed several models within the new Med-Gemini family, a series of models built upon the multimodal foundation of Gemini and fine-tuned on a diverse range of medical data including radiology, histopathology, ophthalmology, dermatology and genomics. We assessed our Med-Gemini models' performance using a comprehensive medical benchmarking suite, including both established benchmarks and custom benchmarks designed to reflect clinical relevance. Notably, some benchmarks involved evaluations by medical experts for tasks such as generating CXR and CT reports and radiology VQA.

Med-Gemini-2D sets a new standard for expert-evaluated chest X-ray report generation, outperforming previous models, and Med-Gemini-3D showcases the first LMM-based report generation for 3D CT. Beyond report generation, Med-Gemini-2D demonstrates exceptional performance in VQA and classification across various medical imaging modalities. Beyond imaging, Med-Gemini-Polygenic outperforms conventional polygenic risk score methods in predicting disease risk. These results demonstrate the potential of the Gemini foundation and the fine-tuned Med-Gemini family in the medical domain. Nonetheless, the results also underscore the need for further rigorous research to ensure safe and effective implementation in real-world clinical settings.

While advanced capabilities on individual medical tasks are useful in their own right, we envision a future in which all of these capabilities are integrated together into comprehensive systems to perform a range of complex multidisciplinary clinical tasks, working alongside humans to maximize clinical efficacy and improve patient outcomes. The results presented in this study represent a step towards realizing this vision.

9. Contributions and Acknowledgments

Contributions

Authors are listed here associated with their primary workstreams. Many authors contributed to additional workstreams beyond the one under which they are listed.

Technical Leads

Lin Yang*,¹
Shawn Xu*,¹
Andrew Selligren*,¹

* Equal contribution

Chest X-Ray

Timo Kohlberger¹
Ira Ktena²
Kendall Park¹
Ryutaro Tanno²
David G. T. Barrett²
Wei-Hung Weng¹
Khaled Saab¹
Tao Tu¹

Computed Tomography

Atilla Kiraly¹
Akshay Goel¹
Arnav Agharwal¹
Nick George¹

Genomics

Cory Y. McLean¹
Yuchen Zhou¹
Farhad Hormozdiari¹
Eric Wang²

Pathology

Dave Steiner¹
Faruk Ahmed¹
Tiam Jaroensri¹
Ellery Wulczyn¹

Additional Contributions

Fayaz Jamil¹
Theo Guidroz¹
Yang Wang³
Siyuan Qiao²
Yun Liu¹
S. Sara Mahdavi²

Clinical

Charles Lau⁴
Sreenivasa Raju Kalidindi⁵
Mozziyar Etemadi⁶
Jorge Cuadros⁷
Gregory Sorensen⁸

Google Research and Google DeepMind Leadership

Shruthi Prabhakara¹
Daniel Tse¹
Shravya Shetty¹
Greg Corrado¹
Katherine Chou¹
Yossi Matias¹
S. M. Ali Eslami²
David Fleet²
Joelle Barral²

Research Leads

Daniel Golden^{†,1}
Shekoofeh Azizi^{†,2}
Rory Pilgrim¹
Christopher Kelly¹

† Equal contribution

¹Google Research

²Google DeepMind

³Verily Life Sciences

⁴Google Research via Advanced Clinical

⁵Apollo Radiology International

⁶Northwestern Medicine

⁷EyePACS, Inc and Meredith Morgan University Eye Center, University of California at Berkeley

⁸DeepHealth / RadNet

Acknowledgements

This project was an extensive collaboration between many teams at Google Research and Google DeepMind. We thank Kevin Swersky and Mike Schaeckermann for their feedback and insight, which significantly contributed to the enhancement of this report. We also thank Sami Lachgar, Lauren Winer, Maggie Shiels, Jessica Valdez, Jon Small, Aaron Abood, Rishad Patel, Christian Wright, Annisah Um'rani, Jean-baptiste Alayrac, Aishwarya Kamath, Viorica Patraucean, Rory Sayres, Abbi Ward, Louis Blankemeier, Olga Kanzheleva, Taedong Yun, Ksenia Konyushkova, Christos Kaplanis, Juanma Zambrano Chaves, Alan Karthikesalingam, Vivek Natarajan, and Can Kirmizi for their valuable insights, technical support and feedback during our research. We thank Kimberly Kanada and Ilana Traynis for their review of the qualitative examples shown in this manuscript. We are grateful to Jonathon Shlens, Dale Webster and Oriol Vinyals for their support during the course of this project. We also thank Michael Colligan and Brittany Stein from DeepHealth/RadNet for their support with data curation.

This research was conducted using the UK Biobank Resource under application number 65275. The results shown here are in part based upon data generated by the [TCGA Research Network](#). The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

Data Availability

Except IND1, CXR-US2, and CT-US1, Eyepacs, and TTH, which are private datasets, the rest of the datasets utilized for developing, benchmarking, and evaluation of Gemini and Med-Gemini in this report are publicly accessible with appropriate permissions. We intend to publicly release our updated classification labels and custom VQA question and answer pairs for the MIMIC-CXR dataset, our splits for the PAD-UFES-20 and VQA-Rad datasets, and several suggested replacement question and answer pairs for the VQA-Rad dataset which were recommended by our reading radiologist. This text will be updated when that data is available.

Code Availability

We will not open-source the model code and weights because of the safety concerns associated with unmonitored use in medical settings. To ensure responsible innovation, we will collaborate with our research partners and healthcare providers to validate and explore safe applications of the Gemini and Med-Gemini through Google Cloud APIs.

Competing Interests

This study was funded by Alphabet Inc and/or a subsidiary thereof ('Alphabet'). Authors who are affiliated with Google Research, Google DeepMind, and Verily Life Sciences are employees of Alphabet and may own stock as part of the standard compensation package.

Use of AI in Manuscript Preparation

This manuscript was written manually, with a small number of copy edits performed using Gemini. The authors take all responsibility for the contents.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.
- Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, et al. Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency. *JAMA Network Open*, 5(8):e2229289–e2229289, 2022.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Clara Albiñana, Zhihong Zhu, Andrew J. Schork, Andrés Ingason, Hugues Aschard, Isabell Brikell, Cynthia M. Bulik, Liselotte V. Petersen, Esben Agerbo, Jakob Grove, et al. Multi-pgs enhances polygenic prediction by combining 937 polygenic scores. *Nature Communications*, 14(4702), 2023.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0447-x. URL <https://doi.org/10.1038/s41591-019-0447-x>.
- Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.
- Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.
- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 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, 2023.
- Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Daniel Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, et al. Pathways: Asynchronous distributed dataflow for ml. *Proceedings of Machine Learning and Systems*, 4:430–449, 2022.
- Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

- Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer, 2023.
- Asma Ben Abacha, Mourad Sarrouiti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania, September 21-24 2021. CEUR-WS.org.
- Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. Stanford crfm introduces pubmedgpt 2.7b. <https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b>, 2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726): 203–209, October 2018.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLi: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.
- Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O’Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):2759–2772, 2020.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- Justin Cosentino, Babak Behsaz, Babak Alipanahi, Zachary R McCaw, Davin Hill, Tae-Hwi Schwantes-An, Dongbing Lai, Andrew Carroll, Brian D Hobbs, Michael H Cho, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirometry identifies new genetic loci and improves risk models. *Nature Genetics*, 55(5):787–795, 2023.
- Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

- 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 *Machine Learning for Health*, pages 209–219. PMLR, 2021.
- Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv preprint arXiv:2404.04667*, 2024.
- Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv preprint arXiv:2308.14089*, 2023.
- Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 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.
- Google. Google’s foundation model for Dermatology. <https://github.com/Google-Health/imaging-research/tree/master/derm-foundation>, 2024. Accessed April 18, 2024.
- Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Shivanand Gornale and Pooja Patravali. Digital knee x-ray images. *Mendeley Data*, 1, 2020.
- Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*, 2024.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv preprint arXiv:2003.10286*, 2020.
- Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.
- Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.
- Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
- Ronnachai Jaroensri, Ellery Wulczyn, Narayan Hegde, Trissia Brown, Isabelle Flament-Auvigne, Fraser Tan, Yuannan Cai, Kunal Nagpal, Emad A Rakha, David J Dabbs, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. *NPJ breast cancer*, 8(1): 113, 2022.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- A Johnson, T Pollard, R Mark, S Berkowitz, and S Horng. MIMIC-CXR database (version 2.0. 0). PhysioNet, 2019a.
- Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg - chest radiographs with structured labels, November 2019b. URL <https://doi.org/10.13026/8360-t248>.
- Alistair EW 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, 2019c.
- Anthony P Khawaja, Jessica N Cooke Bailey, Nicholas J Wareham, Robert A Scott, Mark Simcoe, Robert P Igo Jr, Yeunjo E Song, Robert Wojciechowski, Ching-Yu Cheng, Peng T Khaw, et al. Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nature Genetics*, 50(6):778–782, 2018.
- Atilla P. Kiraly, Corbin A. Cunningham, Ryan Najafi, Zaid Nabulsi, Jie Yang, Charles Lau, Joseph R. Ledsam, Wenxing Ye, Diego Ardila, Scott M. McKinney, Rory Pilgrim, Yun Liu, Hiroaki Saito, Yasuteru Shimamura, Mozziyar Etemadi, David Melnick, Sunny Jansen, Greg S. Corrado, Lily Peng, Daniel Tse, Shravya Shetty, Shruthi Prabhakara, David P. Naidich, Neeral Beladia, and Krish Eswaran. Assistive ai in lung cancer screening: A retrospective multinational study in the united states and japan. *Radiology: Artificial Intelligence*, 2024. doi: 10.1148/ryai.230079. URL <https://pubs.rsna.org/doi/10.1148/ryai.230079>.
- Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018.
- Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.
- Jeremy Lai, Faruk Ahmed, Supriya Vijay, Tiam Jaroensri, Jessica Loo, Saurabh Vyawahare, Saloni Agarwal, Fayaz Jamil, Yossi Matias, Greg S Corrado, et al. Domain-specific optimization and diverse evaluation of self-supervised models for histopathology. *arXiv preprint arXiv:2310.13259*, 2023.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

- Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022.
- 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. *Advances in Neural Information Processing Systems*, 36, 2024.
- Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023a.
- Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023b.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.
- Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, Haixing Dai, Lin Zhao, Lichao Sun, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Xiang Li, Quanzheng Li, and Tianming Liu. Radiology-gpt: A large language model for radiology, 2024b.
- Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, pages 1–12, 2024.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.
- Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.
- Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.
- 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, 2023a.

- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023b.
- Zaid Nabulsi, Andrew Sellergren, Shahar Jamshy, Charles Lau, Edward Santos, Atilla P Kiraly, Wenxing Ye, Jie Yang, Rory Pilgrim, Sahar Kazemzadeh, et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. *Scientific reports*, 11(1):15523, 2021.
- Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):48, 2019.
- Kunal Nagpal, Davis Foote, Fraser Tan, Yun Liu, Po-Hsuan Cameron Chen, David F Steiner, Naren Manoj, Niels Olson, Jenny L Smith, Arash Mohtashamian, et al. Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. *JAMA oncology*, 6(9):1372–1380, 2020.
- Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.
- NHS. Rcr response to nhse data release on diagnostic imaging times, 2024. URL <https://www.rcr.ac.uk/news-policy/latest-updates/rcr-response-to-nhse-data-release-on-diagnostic-imaging-times/>.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023.
- NLST. National Lung Screening Trial (NLST). <https://www.cancer.gov/types/lung/research/nlst>, 2014. Accessed September 15, 2021.
- OpenAI. GPT-4V(ision) Technical Work and Authors. Technical report, OpenAI, 2023. URL <https://cdn.openai.com/contributions/gpt-4v.pdf>.
- Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.
- Ankit Pal and Malaikannan Sankarasubbu. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. *arXiv preprint arXiv:2402.07023*, 2024.
- Pan-UKB team. <https://pan.ukbb.broadinstitute.org>, 2020.
- 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, 2002.
- Kendall Park, Charles Lau, Timo Kohlberger, Tom Pollard, Andrew Sellergren, Rory Sayres, and Atilla P. Kiraly. MIMIC-CXR-GT Database (version 1.0.0), 2024. In preparation.

- Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.
- Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.
- Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR, 2023.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.
- Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*, 2024.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.
- Apaar Sadhwani, Huang-Wei Chang, Ali Behrooz, Trissia Brown, Isabelle Auvigne-Flament, Hardik Patel, Robert Findlater, Vanessa Velez, Fraser Tan, Kamilla Tekiela, et al. Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. *Scientific reports*, 11(1):16605, 2021.
- Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.
- Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.
- Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, March 2015.
- R Summers. Nih chest x-ray dataset of 14 common thorax disease categories. *NIH Clinical Center: Bethesda, MD, USA*, 2019.

- Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 (5), pages 5034–5042, 2024.
- 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.
- Ryutaro Tanno, David Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. *arXiv preprint arXiv:2311.18260*, 2024.
- Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.
- Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Buu Truong, Leland E. Hull, Yunfeng Ruan, Qin Qin Huang, Whitney Hornsby, Hilary Martin, David A. van Hell, Ying Wang, Alicia R. Martin, S. Hong Lee, and Prageep Natarajan. Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics*, 4, 2024.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3):A10a2300138, 2024.
- Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip HS Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*, 2024.
- Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–736. Springer, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

- Felix Vogel, Nina Shvetsova, Leonid Karlinsky, and Hilde Kuehne. VI-taboo: An analysis of attribute-based zero-shot capabilities of vision-language models. *arXiv preprint arXiv:2209.06103*, 2022.
- 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*, pages 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*, pages 11558–11567, 2023a.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *arXiv preprint arXiv:2309.09812*, 2023b.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. Multimodal multitask representation learning for pathology biobank metadata prediction. *CoRR*, abs/1909.07846, 2019.
- Wei-Hung Weng, Andrew Sellergen, Atilla P Kiraly, Alexander D’Amour, Jungyeon Park, Rory Pilgrim, Stephen Pfohl, Charles Lau, Vivek Natarajan, Shekoofeh Azizi, et al. An intentional approach to managing bias in general purpose embedding models. *The Lancet Digital Health*, 6(2):e126–e130, 2024.
- Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. Zero-shot clinical trial patient matching with llms. *arXiv preprint arXiv:2402.05125*, 2024.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.
- Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ digital medicine*, 4(1):71, 2021.
- Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. ELIXR: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.
- An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-nan Hsu. Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015, 2021.
- Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 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, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Yuchen Zhou, Justin T Cosentino, Taedong Yun, Mahantesh I Biradar, Jacqueline Shreibati, Dongbing Lai, Tae-Hwi Schwantes-An, Robert Luben, Zachary R McCaw, Jorgen Engmann, et al. Utilizing multimodal ai to improve genetic analyses of cardiovascular traits. *medRxiv*, pages 2024–03, 2024.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. Protllm: An interleaved protein-language llm with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920*, 2024.

Appendix

A.1. Additional data details

A.1.1. Revised MIMIC-CXR classification labels

One of the limitations of the MIMIC-CXR dataset is the lack of ground-truth labels. MIMIC-CXR JPG (Johnson et al., 2019b) extracted structured labels from 277,827 radiology reports using CheXpert (Irvin et al., 2019), a natural language processing (NLP) tool to extract observations from radiology reports. In order to improve upon these labels on the test subset, we utilized Med-PaLM 2 (Singhal et al., 2023b) coupled with US-based board certified radiologists to refine those labels. This work is further adjudication of the labels used in Xu et al. (2023). We first used a keyword search to identify reports containing text associated with the a given finding (e.g., “Cardiomegaly”). Next, Med-PaLM 2 was applied to the flagged radiology reports on a per-label basis using two queries shown in Table A.1 for a total of 23,824 queries. All identified positive and negative labels that disagreed with the original labels were flagged for human verification.

Three US-based board certified radiologists reviewed the 1,378 flagged labels and a fourth academic US-based board certified thoracic radiologist adjudicated the responses of reviewer disagreements. For each finding and report, radiologists selected one of four possible labels defined by MIMIC-CXR JPG: positive, negative, uncertain, and not mentioned. Zero-round adjudication was performed on the reviewers’ annotations. There was strong inter-rater agreement (Fleiss’ $\kappa = 0.71$); reviewers were unanimous for 77% of the labels. In cases of disagreement between reviewers, majority vote was used (21%), and when all three reviewers disagreed (2%) a senior academic thoracic radiologist provided the final determination.

In the final analysis of the flagged reports and findings, Med-PaLM 2’s label matched the ground truth 66% of the time while the original labels were correct in 19% of the cases. The labels are in preparation to be released (Park et al., 2024).

Table A.1 | Prompt styles used for Med-Palm2 in extracting structured labels. Two prompt styles were used to extract labels from the given MIMIC-CXR radiology reports. The following shows the template used for identifying fractures.

| Prompt Style | Prompt Example |
|--------------|---|
| Bot | <p>[Bot] I’m a helpful radiology assistant, who provides concise answers to questions about information in a chest x-ray report.</p> <p>[User] Determine the answer to the following question: [Does the patient have a fracture?], given the context of the follow chest x-ray report: <REPORT TEXT> Do not mention conditions or parts of the report not relevant to the question. Make sure to only answer: [Does the patient have a fracture?]</p> <p>[Bot]</p> |
| Question | <p>You are a helpful medical knowledge assistant. Provide useful, complete, concise, and scientifically-grounded queries to radiology reports.</p> <p>Does this report mention that the patient has a fracture? Report: <REPORT TEXT></p> |

A.1.2. Prompts for VQA and CXR classification evaluations

We explored both binary question prompting for each of the evaluated top 5 conditions for MIMIC-CXR including the abnormal/normal class, as well as multi-select prompts for all of them at once on the validation set. Since binary prompts overall yielded better macro F1 scores for Med-Gemini, these were used at evaluation. For Gemini Ultra binary question were used for the top-5 conditions and a multi-select prompt for the normal/abnormal condition, see Tables A.3 and A.4. Each prompt template was crafted for each model in aiming to optimize its performance, which yielded very short prompt templates for both classification and VQA for Med-Gemini, since it was fine-tuned with clinical questions.

Table A.2 | Zero-shot prompt templates used for VQA evaluations (excluding formatting tokens). The following prompt templates were used for each model and VQA evaluation data set. The < question > placeholder was replaced with the individual question of each VQA triplet.

| VQA Dataset | Model | Prompt Template |
|-------------|------------|---|
| VQA-Rad | Gemini | < image > You are a helpful radiology assistant. Given this radiology image, which can be frontal chest X-ray provide a very short, concise answer, like: “pleural effusion,” “yes,” “right,” to the following question: < question > |
| | Med-Gemini | < image > < question > |
| Slake-VQA | Gemini | < image > You are a helpful radiology assistant. Given this radiology image, which can be a frontal chest X-ray, a single slice head or abdominal CT or MR image, provide a very short, concise answer, like: “pleural effusion,” “yes,” “right,” to the following question: < question > |
| | Med-Gemini | < image > < question > |
| PathVQA | Gemini | < image > You are a helpful pathology assistant. Given the pathology image, provide a concise answer, like: “pleural effusion,” “yes,” “right,” to the following question: < question > |
| | Med-Gemini | < image > < question > |

A.1.3. New balanced splits for VQA-Rad dataset

The official train/test split of the VQA-Rad (Lau et al., 2018) dataset comprises 1,797 QA pairs for training (i.e. dataset field QID_para ∈ {‘freeform’, ‘para’}) for 313 different MedPIX@images (per field IMAGEID), and 451 QA pairs for 203 different images in the test set (i.e. field QID_para ∈ {‘test_freeform’, ‘test_para’}). Although images were sampled such that each is not only for a different case, but also a different patient, see (Lau et al., 2018), 202 of the test IMAGEIDs and also match the train set IMAGEIDs. Hence most of the test images also appear in the train set, only the questions and answers differ. For some VQAs even the latter is not completely true, since VQA-Rad contains paraphrased questions which share the same answer.

To remove this train/test contamination issue, in Xu et al. (2023) we proposed a different validation/test split, which is based on the IMAGEIDs in order to ensure disjoint images and thus patients. In this work we split this relatively large test set further into a new test and train set, which are roughly equal-sized, and assign a few remaining former test set IMAGEIDs and corresponding VQAs to the existing validation set, such that all three new splits not only are roughly equal-sized, but their ratio of open to closed questions (as determined by field A_TYPE) are approximately equal within each of the three depicted anatomical regions (chest, head and abdomen), see Table A.5. We chose to equalize this ratio since open-ended questions are more difficult for AI models, and the level of difficulty ought to be similar for each new split. Similarly, while swapping individual IMAGEIDs

Table A.3 | Zero-shot prompts used evaluating classification performance on chest X-ray. Binary questions were employed for Med-Gemini on all CXR classification conditions, and for Gemini Ultra on all except for the normal/abnormal condition (label “No Finding”). The < question > placeholder was replaced with the label-dependent text listed in Table A.4, while < view position > with the respective MIMIC-CXR meta information (e.g. “AP”).

| Model | Used for | Prompt template |
|------------|--------------------------------------|--|
| Gemini | Top-5 Conditions | <p>< image > You are a helpful radiology assistant.</p> <p>The following is a question about findings in chest X-ray in the frontal view. Answer only with Yes or No.</p> <p>Q: Given the X-ray < view position > image, < question > indicated in this chest X-ray image?</p> <p>A:</p> |
| Med-Gemini | Top-5 Conditions and Normal/Abnormal | < image > < question > |
| Gemini | Normal/Abnormal | <p>< image > You are a helpful radiology assistant.</p> <p>The following are multiple-choice questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray by responding with the corresponding answer choice letter(s).</p> <p>Q: Given the X-ray image, which of the following abnormalities are indicated by the image?</p> <p>(A) Atelectasis (B) Cardiomegaly (C) Consolidation (D) Edema (I) Pleural Effusion (O) No Abnormality (E) Enlarged Cardiome-diastinum (F) Fracture (G) Lung Lesion (H) Lung Opacity (I) Pleural Effusion (J) Pleural Other (K) Pneumonia (L) Pneumothorax (M) Support Devices</p> |

and associated question and answers, we approximately equalized the distribution of question types (field Q_TYPE) within each split, in order to gain similar ones, see Table A.6.

A.1.4. Polygenic risk prediction

We crafted prompts for predicting the status of various health outcomes using both an individual’s PRS image and their demographic information. An example prompt for predicting coronary artery disease is shown in Table A.9.

For linear probes of out-of-distribution outcomes, we used data of a related in-distribution outcome to train the linear probe and then evaluated the predictions on the out-of-distribution outcome. For example, in order to evaluate diabetic retinopathy, we train the linear probe to predict type 2 diabetes and evaluate the type 2 diabetes predictions on diabetic retinopathy data. In general, the most related in-distribution outcome is defined as the outcome with the highest Matthew’s correlation coefficient with the out-of-distribution outcome across individuals in our training set (Table A.11). To evaluate Med-Gemini-Polygenic, we directly prompted Med-Gemini-Polygenic to predict the out-of-distribution outcome without providing any information about correlations between out-of-distribution and in-distribution outcomes.

Table A.4 | Question arguments used for binary classification prompt templates. We used the following arguments for the < question > placeholder in the binary classification prompts, Table A.3, which were triggered depending on the corresponding CheXpert label to be in {0.0, 1.0}.

| Condition | Argument for Med-Gemini prompt | Argument for Gemini prompt |
|------------------|--------------------------------|----------------------------|
| Atelectasis | Is there atelectasis? | is atelectasis |
| Cardiomegaly | Is there cardiomegaly? | is cardiomegaly |
| Consolidation | Is there consolidation? | is consolidation |
| Edema | Is there pulmonary edema? | is pulmonary edema |
| Pleural Effusion | Is there pleural effusion? | is pleural effusion |

Table A.5 | Distribution of VQA-Rad answer types (closed- vs. open-ended) across the new balanced splits for each anatomical region. The new train/validation/test splits not only guarantee the images and thus patients to be disjoint, but also provide a similar ratio of open to close-ended questions (dataset field A_TYPE) per anatomical region, which typically affects VQA performance of AI models due to open-ended questions being more difficult to answer correctly.

| Anatomical Region | Split (balanced) | Number of question-answers pairs | | | Ratio of open to closed |
|-------------------|------------------|----------------------------------|--------|-------|-------------------------|
| | | Open | Closed | Total | |
| Abdomen | Train | 104 | 153 | 257 | 68.0% |
| | Validation | 97 | 143 | 240 | 67.8% |
| | Test | 99 | 143 | 242 | 69.2% |
| Chest | Train | 94 | 185 | 279 | 50.8% |
| | Validation | 107 | 161 | 268 | 66.5% |
| | Test | 82 | 165 | 247 | 49.7% |
| Head | Train | 131 | 118 | 249 | 111.0% |
| | Validation | 113 | 107 | 220 | 105.6% |
| | Test | 122 | 144 | 246 | 84.7% |
| All | | 949 | 1,299 | 2,248 | 73.1% |

Table A.6 | Distribution of VQA-Rad question types across the new balanced splits. In addition to comparable open-to-closed-question ratios, the question types (dataset field Q_TYPE) distribute similarly as well.

| Anatomical Region | Abdomen | | | Chest | | | Head | | | All |
|-------------------|---------|------|------|-------|------|------|-------|------|------|------|
| | Train | Val. | Test | Train | Val. | Test | Train | Val. | Test | |
| Total | 257 | 240 | 242 | 279 | 268 | 247 | 249 | 220 | 246 | 2248 |
| Presence | 105 | 95 | 100 | 101 | 92 | 93 | 79 | 67 | 68 | 800 |
| Positional | 26 | 27 | 21 | 37 | 36 | 42 | 35 | 42 | 50 | 316 |
| Abnormality | 19 | 18 | 27 | 15 | 40 | 21 | 22 | 25 | 15 | 202 |
| Other | 20 | 20 | 22 | 26 | 19 | 19 | 22 | 14 | 32 | 194 |
| Modality | 25 | 13 | 23 | 19 | 17 | 13 | 35 | 24 | 16 | 185 |
| Size | 17 | 20 | 11 | 41 | 26 | 27 | 10 | 5 | 14 | 171 |
| Plane | 12 | 9 | 13 | 14 | 12 | 13 | 17 | 16 | 14 | 120 |
| Attribute Other | 13 | 19 | 7 | 5 | 9 | 8 | 4 | 12 | 10 | 87 |
| Organ | 5 | 4 | 2 | 10 | 8 | 7 | 13 | 3 | 7 | 59 |
| Color | 6 | 7 | 8 | 2 | 0 | 0 | 9 | 7 | 13 | 52 |
| Counting | 3 | 8 | 3 | 5 | 1 | 2 | 0 | 1 | 1 | 24 |

Table A.7 | Examples of some of our curated captions for histopathology patches used in our training set.

| Task | Curated Captions |
|--|---|
| Breast cancer detection in lymph nodes | Microscopic view of a lymph node with infiltrating malignant epithelial cells consistent with metastatic breast carcinoma. |
| | Lymph node section on histopathology demonstrating features of metastatic carcinoma, likely originating from the breast. |
| | Region of a lymph node on histopathology, showing predominantly lymphocytes and immune cells. |
| Histologic subtyping for lung adenocarcinoma | H&E histopathology image demonstrating acinar subtype lung adenocarcinoma with glandular formation. |
| | Cribriform pattern in lung adenocarcinoma with closely packed, back-to-back glands lacking an organized pattern. |
| | Microscopic view of lung adenocarcinoma showing tumor cells spreading along the preexisting alveolar architecture, consistent with the lepidic subtype. |
| Prostate Gleason grading | Microscopic view of a prostate biopsy with discrete, uniformly sized and shaped glands consistent with Gleason pattern 3 prostate cancer. |
| | Prostate biopsy image highlighting areas of fused glands or poorly formed glands, consistent with Gleason pattern 4 carcinoma. |
| | Microscopic view of a prostate tissue with discrete, uniformly sized and shaped glands consistent with Gleason pattern 3 prostate cancer. |
| Breast cancer nuclear pleomorphism grading | Microscopic view of invasive breast carcinoma showing moderate nuclear atypia, including enlarged nuclei and prominent nucleoli (nucleoleomorphism score 2). |
| | Focus on infiltrating tumor cells within an H&E stained image, demonstrating bland nuclear features suggestive of low-grade invasive breast carcinoma. |
| | H&E histopathology image demonstrating invasive breast carcinoma with high-grade nuclear features (nucleoleomorphism score 3). |
| Breast cancer tubule formation grading | This microscopic view demonstrates a predominance of well-defined glandular structures, indicating a tubule formation score of 1 in this breast carcinoma. |
| | The presence of some discernible tubules, alongside regions of less-defined glandular architecture, indicates a tubule formation score of 2. |
| | A tubule formation score of 3 is evident within this H&E stained image, where the invasive carcinoma shows a scarcity of well-defined glandular structures. |
| Cervical dysplasia grading | H&E stained image of a cervical biopsy demonstrating features of cervical intraepithelial neoplasia grade 1 (CIN 1), including nuclear atypia in the lower third of the epithelium. |
| | Microscopic view of a cervical biopsy with CIN 1, showing enlarged nuclei and increased nuclear-to-cytoplasmic ratio in the basal layer of the epithelium. |
| | H&E histopathology of a cervical biopsy focusing on high-grade dysplasia, showing prominent nuclear abnormalities and disruption of the normal epithelial architecture. |

Table A.8 | Overview of patch-level histopathology datasets used for fine-tuning and linear-probe evaluation. Tasks adopted from (Lai et al., 2023). The OOD column indicates whether the task was included in Med-Gemini fine-tuning. Number of slides shows the counts split across train, validation, and test sets, respectively.

| Dataset | Task | Number of slides | Classes | OOD |
|---------------------------------------|--|------------------|----------------------------------|-----|
| CAMELYON16 (Bejnordi et al., 2017) | Breast cancer detection in lymph nodes | 216/54/258 | Tumor, Non-Tumor | - |
| Lung AD (Sadhvani et al., 2021) | Histologic subtyping for lung adenocarcinoma | 73/25/50 | 9 classes* | - |
| Gleason NCB (Nagpal et al., 2020) | Gleason grading on prostate needle core biopsies | 178/85/88 | Benign, GP3, GP4, GP5 | - |
| Gleason RP (Nagpal et al., 2019) | Gleason grading on radical prostatectomies | 550/259/202 | Benign, GP3, GP4, GP5 | - |
| Breast IC (Jaroensri et al., 2022) | Breast invasive carcinoma detection | 573/288/669 | Benign, DCIS, Invasive Carcinoma | - |
| Breast NP (Jaroensri et al., 2022) | Breast cancer nuclear pleomorphism grading | 681/343/945 | NP1, NP2, NP3 | - |
| Breast TF (Jaroensri et al., 2022) | Breast cancer tubule formation grading | 681/343/945 | TF1, TF2, TF3 | - |
| CIN (Lai et al., 2023) | Cervical dysplasia grading | 329/74/229 | Non-tumor, CIN 1, CIN 2+ | - |
| CRC (Wulczyn et al., 2021) | Colorectal carcinoma detection | 149/51/44 | Tumor, Non-Tumor | - |
| Tissue type (Weng et al., 2019) | Tissue type classification (internal dataset) | 17319/6488/6719 | 16 tissue types [†] | ✓ |
| TCGA study (Lai et al., 2023) | TCGA study type classification | 2952/1466/1489 | 10 TCGA study types [‡] | ✓ |

* Lung AD histologic subtypes and other classes: Acinar, Cribriform, Lepidic, Micropapillary, Papillary, Solid, Leukocyte, Necrosis, Non-tumor.

[†] Tissue types: Appendix, Breast, Cervix, Colon and rectum, Fallopian Tube, Gallbladder, Liver, Lymph node, Ovary, Placenta, Prostate, Skin, Thyroid, Upper GI, Uterus, Vas deferens.

[‡] TCGA study types: BLCA, BRCA, COAD, HNSC, KIRC, LIHC, LUAD, LUSC, OV, STAD.

Table A.9 | Examples of a prediction prompt for coronary artery disease using an individual’s PRS image and demographic information. For privacy reasons, the PRS image shown is an average PRS image over 100 individuals.

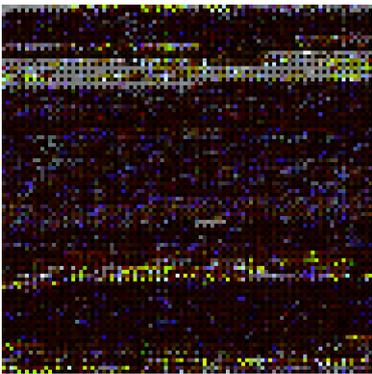
| Image | Prompt | Target |
|---|--|--------|
|  | <p> Given this genomic data, and the following health information:</p> <p>age: 60;</p> <p>sex: female;</p> <p>body mass index: 26.1;</p> <p>predict whether the individual has the following condition or not. Respond “Yes” or “No”.</p> <p>Coronary artery disease:</p> | No |

Table A.10 | Overview of UK Biobank fields used to compile health outcomes for training and evaluating Med-Gemini-Polygenic.

| Health outcome | UK Biobank fields involved | Conversion process |
|-------------------------|---|--|
| Coronary artery disease | 131307 | If source of report of I25 (chronic ischaemic heart disease) available (field 131307). |
| Stroke | 6150, 131368, 131369, 41271, 41270 | Logical OR of: choice 3 of touchscreen questionnaire (field 6150), first reported stroke (field 131368), source of report stroke (field 131369), ICD-9 code 434.91 in field 41271, ICD-10 codes I63.* and I64.* in field 41270. |
| Type 2 Diabetes | 20002, 41270 | Code 1223 in self-reported non-cancer illness code (field 20002), or ICD-10 codes E11.* in field 41270. |
| Glaucoma | 4689, 5326, 5327, 6148, 20003, 41202, 41203, 41205, 42104 | See Khawaja et al. (2018) for details. |
| COPD | 6152, 41270, 41271, 42040 | See Cosentino et al. (2023) for details. |
| Rheumatoid arthritis | 131851 | If source of report of M06 (field 131851). |
| Major depression | 20126, 41270 | Code 3, 4, 5 in bipolar and major depression status (field 20126), or ICD-10 codes F32.* and F33.* in field 41270. |
| All-cause mortality | 40000 | If there is a value in date of death (field 40000). |
| Hypertension | 20002, 41271, 41270 | Logical OR of: Code 1065 in self-reported non-cancer illness code (field 20002), ICD-9 code 401.* and 405.* in field 41271, ICD-10 codes I10 and I15.* in field 41270. |
| Hypercholesterolemia | 30780, 30760 | Labels are calculated using LDL (30780) and HDL(30760) by: LDL cholesterol \geq 190 mg/dL OR (LDL \geq 160 mg/dL AND (HDL $<$ 40 mg/dL AND person is Male) OR (HDL $<$ 55 mg/dL AND person is Female))) Data reported in units of mmol/L, to convert to mg/dL, multiply these by 38.67. |
| Atrial Fibrillation | 20002, 41271, 41270 | Logical OR of: Code 1471 in self-reported non-cancer illness code (field 20002), ICD-9 code 4273 in field 41271, ICD-10 codes I48(0 1 2 9) in field 41270. |
| Diabetic Retinopathy | 6148, 20002, 5890, 5901, 41271, 41270 | Logical OR of: Code 1276 in self-reported non-cancer illness code (field 20002), Code 1 in eye problem/disorder (6148), Code 1, 2, 3 in Which eye(s) affected by diabetes-related eye disease(5890), There is a value in Age when diabetes-related eye disease diagnosed (5901), ICD-9 code 3620 in field 41271, ICD-10 codes H360 in field 41270. |
| Asthma | 20002, 41270 | Code 1111 in self-reported non-cancer illness code (field 20002), or ICD-10 codes J45.*, J46.* in field 41270. |
| Pneumonia | 41271, 41270 | ICD-9 code 401.* and 480, 481, 482, 483, 484, 486 in field 41271, ICD-10 codes J12.*, J13, J14, J15.*, J16.*, J17.*, J18.* in field 41270. |

Table A.11 | Polygenic risk prediction health outcomes correlation. Most correlated In-distribution (ID) health outcomes for each out-of-distribution (OOD) health outcome. In general, the most related in-distribution outcome is defined as the outcome with the highest Matthew’s correlation coefficient with the out-of-distribution outcome across individuals in our training set.

| OOD health outcome | ID health outcome | Matthew’s correlation coefficient |
|---------------------------|--------------------------|--|
| Hypertension | Coronary Artery Disease | 0.266 |
| Hypercholesterolemia | Major Depression | 0.007 |
| Atrial Fibrillation | Coronary Artery Disease | 0.240 |
| Diabetic Retinopathy | Type 2 Diabetes | 0.325 |
| Asthma | COPD | 0.186 |
| Pneumonia | All-cause Mortality | 0.352 |

A.2. Additional results

A.2.1. Data-efficient classification

We performed data-efficient classification for Chest X-ray classification task focusing on examples across 8 different findings (atelectasis, cardiomegaly, airspace opacity, consolidation, fracture, pneumothorax, pleural effusion, and pulmonary edema). We also deploy two out-of-distribution datasets including ChestX-ray14 and CheXpert for this purpose. Our data-efficient classification follows the protocol from (Xu et al., 2023) except that instead of training a Multilayer Perceptron (MLP) as a nonlinear classifier, we train a linear probe on top of the frozen image encoder. Following the ELEVATER (Li et al., 2022) method, we initialize the weights of the final linear layer with the text embeddings for the class label. Training parameters includes a learning rate of 0.2, a batch size of 512, and 300 epochs utilizing the Layer-wise Adaptive Rate Scaling (LARS) optimizer.

In alignment with previous best-in-class method, ELIXR (Xu et al., 2023), the linear classifiers were trained on 5 different varying sample sizes including 0.01% to 100% subsets of the training data to facilitate direct comparability of results to Xu et al. (2023). The smallest sample size includes 64 samples. Figure A.1 shows aggregated results of Med-Gemini vs. ELIXR on ChestX-ray14 and CheXpert (Xu et al., 2023) for 5 and 6 various runs, respectively. Comparison between data-efficient classification results of Med-Gemini vs. ELIXR reveals that linear probes trained on top of visual embeddings from Med-Gemini exhibit robust performance in data-efficient classification, although approximately one order of magnitude inferior than ELIXR at the sample size as low as 64 samples.

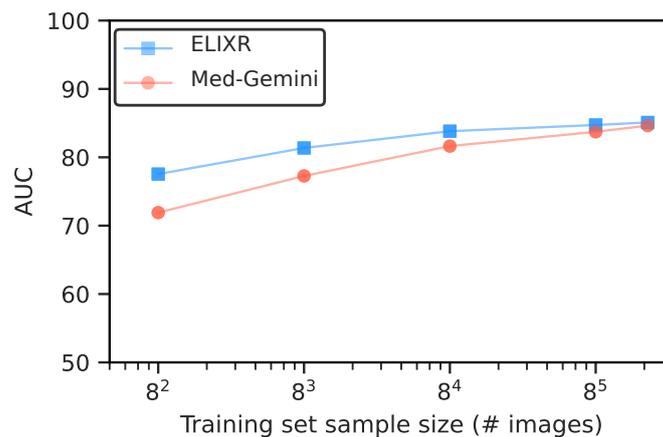


Figure A.1 | Data-efficient classification results for Med-Gemini vs. ELIXR. We target classification across 8 different findings (atelectasis, cardiomegaly, airspace opacity, consolidation, fracture, pneumothorax, effusion, and pulmonary edema) and 2 out-of-distribution datasets (ChestX-ray14 and CheXpert). Linear probes trained on top of visual embeddings from Med-Gemini show strong performance on classification although roughly one order of magnitude inferior to ELIXR at sample sizes as low as 64 examples.

A.2.2. Polygenic risk prediction

Beyond evaluating Med-Gemini-Polygenic, we also compared linear probes of the Med-Gemini-Polygenic embeddings to linear probes of the demographics only and the ensemble of PRSs and demographics. Figure A.2 uses the same balanced sets of 400 individuals as used in Figure 5, and Figure A.3 uses larger balanced sets containing all the positive cases per health outcome and an equal number of controls. The AUC metrics are relatively consistent between both evaluation sets. Furthermore, we computed Med-Gemini-Polygenic performance on coronary artery disease and COPD in 4000-sample evaluations (Figure A.4), and observed stable results. Taken together, these results suggesting that our evaluation set of 400 individuals is representative of overall model performance.

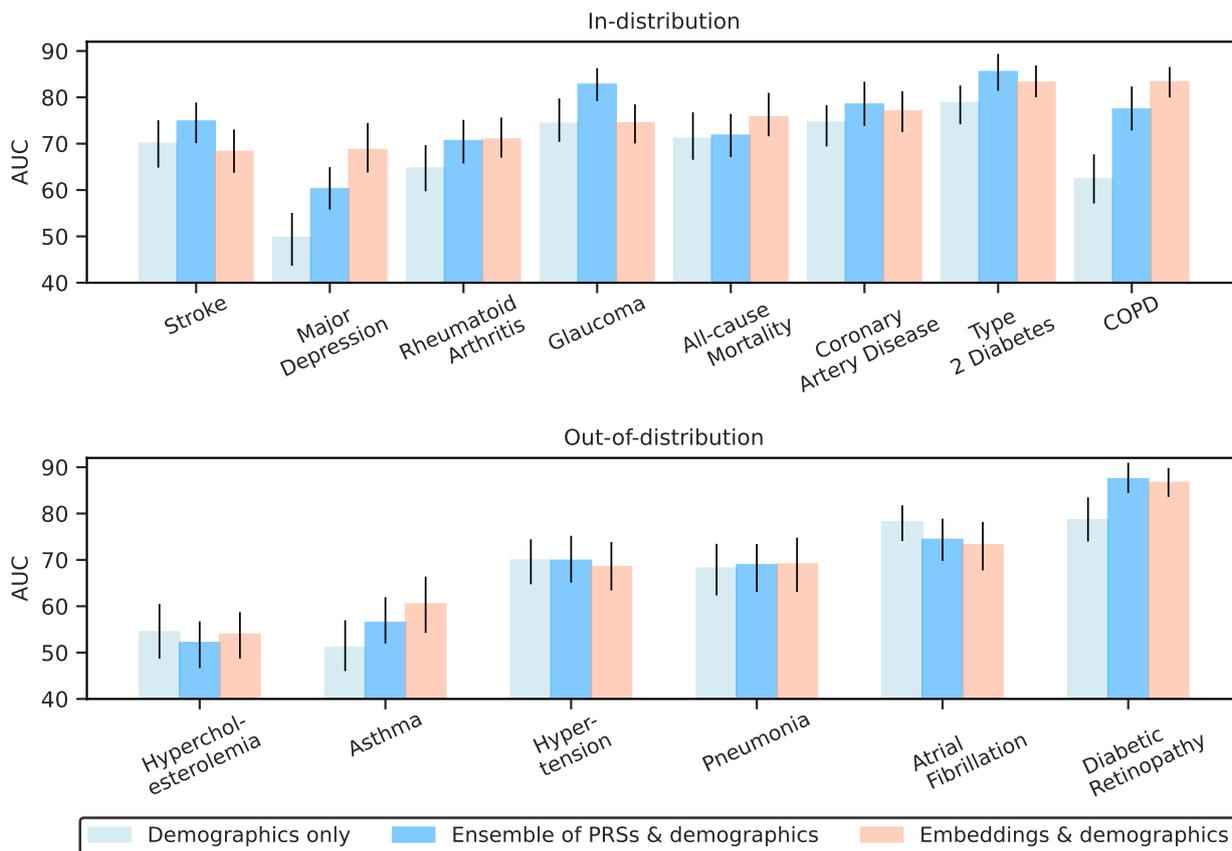


Figure A.2 | Health outcome prediction using linear probes for both in-distribution (ID) and out-of-distribution (OOD) outcomes on balanced evaluation sets of 400 individuals. “Demographics only” used a linear probe of age, sex, and BMI to predict each health outcome, “Ensemble of PRSs and demographics” combined demographics with all 7,145 PRSs in a linear probe, and “Embeddings and demographics” combined demographics with genetic risk embeddings. For OOD health outcomes, the linear probes were trained to predict the most-correlated ID outcome (Table A.11), and those predictions were then evaluated on the OOD outcome.

In addition, we demonstrated that using the Med-Gemini-Polygenic framework likely results in better predictive performance than linear models trained with all PRS featurizations plus demographics regardless of future sample sizes available by conducting sample size ablation tests on the PRS ensemble models. We observed performance plateaus for the linear model with at most 10^4 samples (Figure A.6).

Finally, we investigated the relative contributions of the genomic embedding and modeling non-linear interactions between genomic representations and demographic information by comparing the performance of Med-Gemini-Polygenic to two other non-linear models: a gradient-boosted decision tree (GBDT) of the “Embeddings and demographics” (“Embeddings”) and a GBDT of the most correlated individual PRS at each of the three significance thresholds and demographics (“Best PRSs”). Med-Gemini-Polygenic and the GBDT of “Embeddings and demographics” yield comparable performance across all traits, and consistently outperform the GBDT of “Best PRS” for in-distribution outcomes, confirming the importance of both multi-PRS predictors and accurately modeling non-linear interactions between genetic contributors and demographic information (Table A.12).

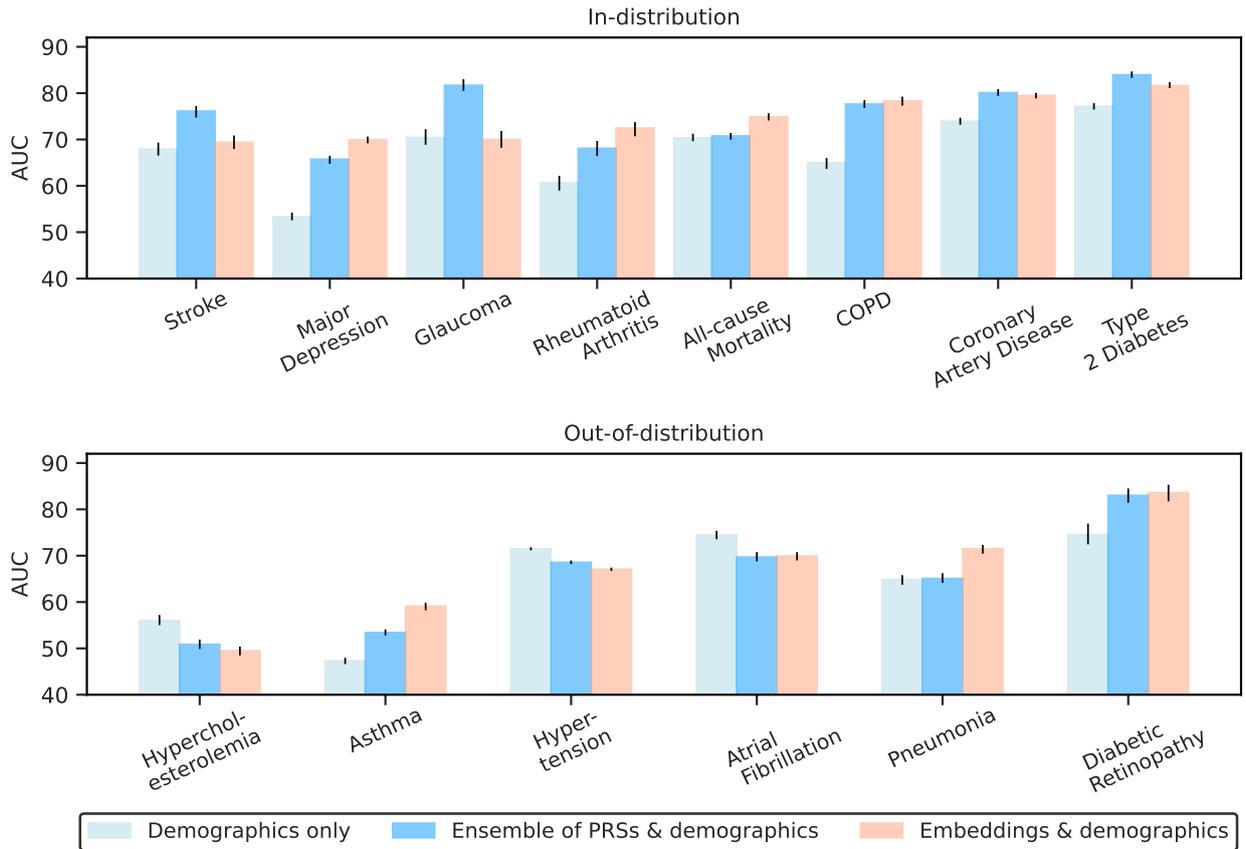


Figure A.3 | Health outcome prediction using linear probes for both in-distribution (ID) and out-of-distribution (OOD) outcomes on larger balanced evaluation sets. For each outcome, the evaluation set includes all positive cases in our test set and the same number of negative controls. “Demographics only” used a linear probe of age, sex, and BMI to predict each health outcome, “Ensemble of PRSs and demographics” combined demographics with all 7,145 PRSs in a linear probe, and “Embeddings and demographics” combined demographics with genetic risk embeddings. For OOD health outcomes, the linear probes were trained to predict the most-correlated ID outcome (Table A.11), and those predictions were then evaluated on the OOD outcome.

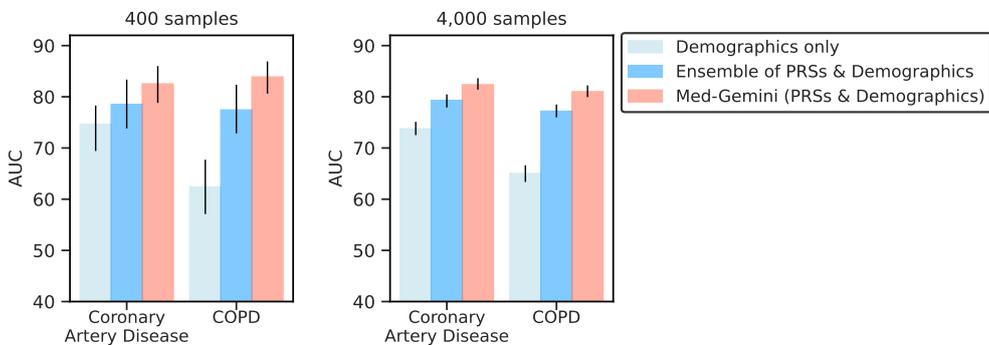


Figure A.4 | Health outcome prediction performance on evaluation sets of 400 and 4,000 balanced samples for coronary artery disease and COPD. The left plot replicates Figure 5 results. The right plot is the analogous performance in larger samples of 4,000 balanced case-control datasets.

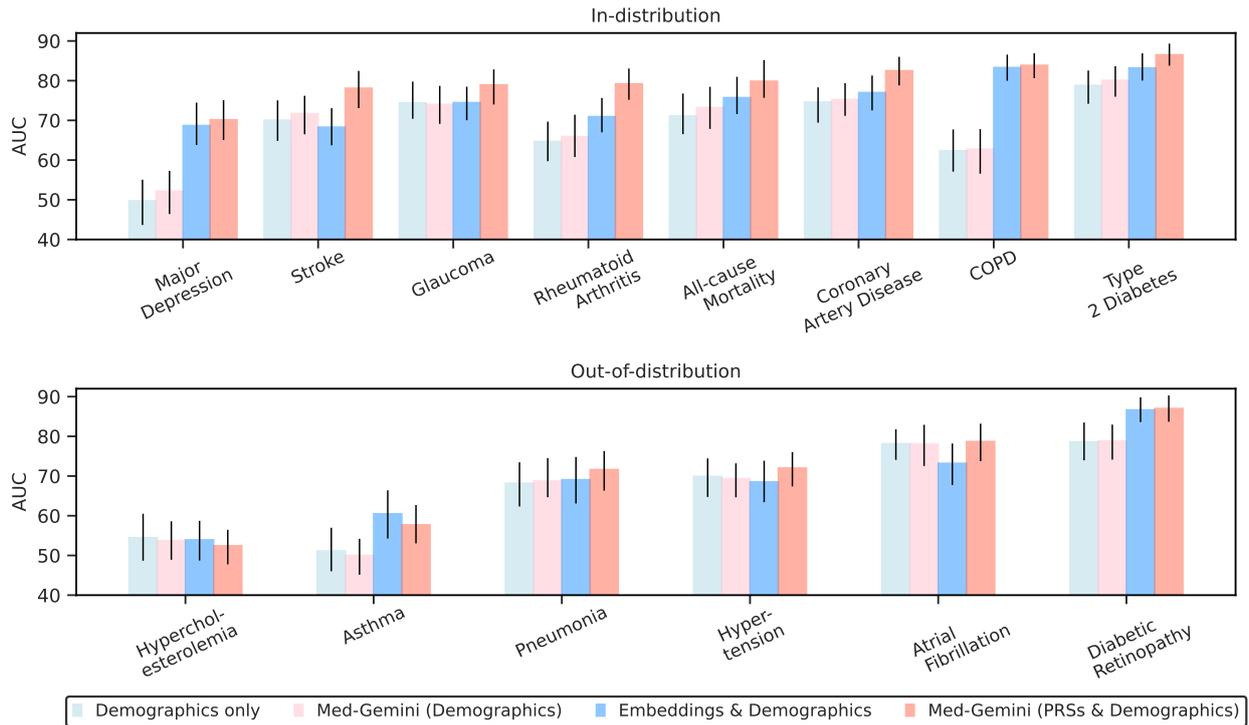


Figure A.5 | Health outcome prediction using linear probes and Med-Gemini-Polygenic for both in-distribution (ID) and out-of-distribution (OOD) outcomes on balanced evaluation sets of 400 individuals. “Demographics only” used a linear probe of age, sex, and BMI to predict each health outcome, and “Embeddings and demographics” combined demographics with genetic risk embeddings. Med-Gemini-Polygenic was prompted with either demographic information only or the individual’s PRS image and demographic information. For OOD health outcomes, the linear probes (“Demographics only” and “Embeddings and demographics”) were trained to predict the most-correlated ID outcome (Table A.11), and those predictions were then evaluated on the OOD outcome.

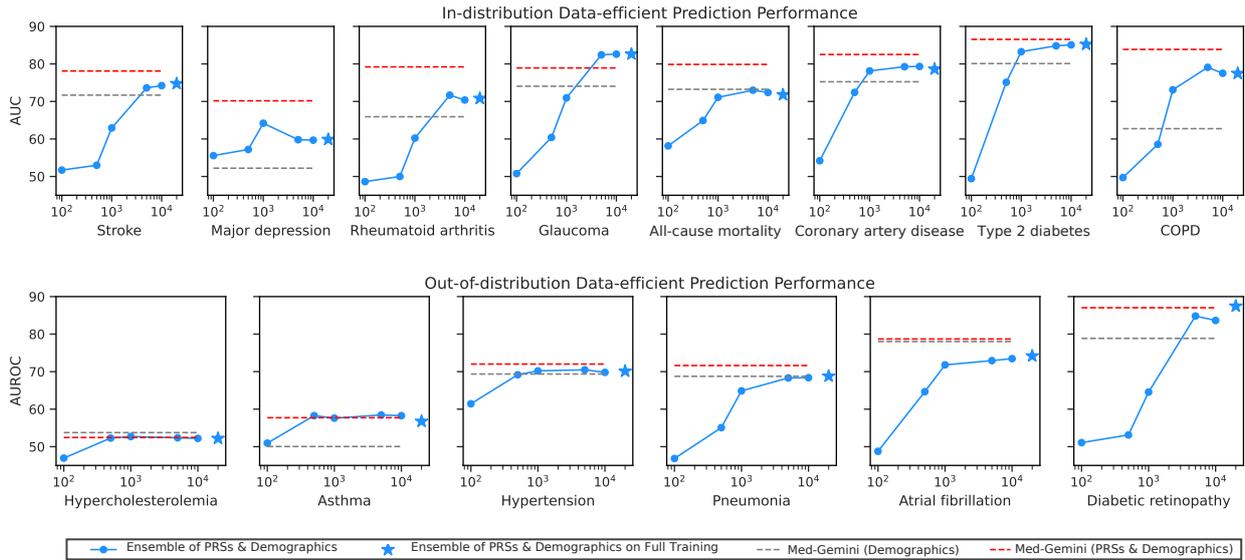


Figure A.6 | Comparing health outcomes prediction performance of Med-Gemini-Polygenic (zero-shot for out-of-distribution outcomes) with PRS plus demographics linear probes trained with different sample sizes. “Ensemble of PRSs & Demographics” are linear models trained against the given health outcome (or the most related ID health outcome for OOD outcomes) using all 7,145 PRSs and demographics. Dashed constant lines show the performance of Med-Gemini-Polygenic on predicting outcomes, given only demographics or genetic risk “image” plus demographics. All linear models in this experiment were trained on population-prevalence data splits.

Table A.12 | Prediction performance measured by AUC for non-linear models of genomics and demographics. The strong performance of Med-Gemini-Polygenic stems from both the inclusion of multiple PRS in the genomic representation and capturing the non-linear interactions between genomics and demographics. “Embeddings”, a GBDT of the Med-Gemini embeddings and demographics. “Best PRS”, a GBDT of the most correlated individual PRS with the outcome at each significance level and demographics.

| Outcome | Med-Gemini-Polygenic | Embeddings | Best PRS |
|-------------------------|----------------------|-------------|-------------|
| Major Depression | 70.2 | 68.3 | 57.1 |
| Stroke | 78.1 | 77.3 | 75.0 |
| Glaucoma | 78.9 | 78.1 | 80.4 |
| Rheumatoid Arthritis | 79.2 | 79.0 | 76.4 |
| All-cause Mortality | 79.9 | 80.3 | 75.3 |
| Coronary Artery Disease | 82.5 | 79.7 | 77.3 |
| COPD | 83.9 | 83.9 | 72.7 |
| Type 2 Diabetes | 86.5 | 86.1 | 85.3 |
| Hypercholesterolemia | 52.4 | 54.1 | 53.5 |
| Asthma | 57.7 | 59.0 | 63.9 |
| Pneumonia | 71.6 | 73.3 | 70.3 |
| Hypertension | 72.0 | 70.2 | 72.1 |
| Atrial Fibrillation | 78.7 | 78.4 | 68.5 |
| Diabetic Retinopathy | 87.0 | 89.0 | 84.5 |

A.2.3. MIMIC-CXR classification

Table A.13 shows the comparison between performance of Med-Gemini and Gemini Ultra measure by F1-score for the original label and revised label as explained in Section A.1.1. Revised MIMIC-CXR labels significantly improve chest X-ray classification performance measured by F1-score (%). Our results demonstrate the impact of accurate ground truth on model evaluation.

Table A.13 | MIMIC-CXR classification performance using the revised labels vs. the original. Performance on chest X-ray MIMIC-CXR classification measured by F1-score (%) when using the original vs. the revised labels. Our experiment shows the measured performance is improved when the revised version of labels are used as the ground-truth.

| Datasets | Condition | Original Labels - F1(%) | | Revised Labels - F1(%) | |
|-----------|------------------------|-------------------------|--------------|------------------------|--------------|
| | | Med-Gemini | Gemini Ultra | Med-Gemini | Gemini Ultra |
| MIMIC-CXR | Atelectasis | 99.3 | 90.7 | 99.8 | 88.1 |
| | Cardiomegaly | 92.8 | 93.8 | 94.1 | 94.6 |
| | Consolidation | 79.5 | 76.3 | 82.0 | 77.0 |
| | Edema | 86.7 | 86.1 | 86.8 | 86.4 |
| | Pleural Effusion | 91.0 | 89.3 | 90.8 | 88.1 |
| | Class-weighted average | 92.1 | 89.0 | 91.2 | 88.6 |
| | Normal/Abnormal | 40.9 | 6.3 | 42.0 | 29.7 |

A.2.4. Histopathology classification

Table A.14 details the linear probing results for the histopathology patch-classification task, reporting 1-vs-rest AUC (%) with 95% confidence intervals. The confidence intervals were obtained using blocked bootstrap resampling over test set slides with 10,000 replicates. Our model’s image embeddings match the performance of the histopathology-specialized model (PathSSL) on 6 out of 9 in-distribution tasks, with room for improvement on the remaining tasks. While Gemini and Med-Gemini-2D perform similarly overall, Med-Gemini shows a trend towards higher mean AUC on most in-distribution tasks and both out-of-distribution tasks.

Table A.14 | Histopathology patch-classification with linear-probe. We measure the 1-*vs*-rest AUCs (%) from linear-probing on our histopathology patch-classification tasks. The 95% confidence intervals were obtained by a blocked bootstrap resampling (over test set slides) using 10000 replicates. The image embeddings from Med-Gemini perform on par with a histopathology-specialized foundation model (PathSSL) on 6 of the 9 in-distribution tasks. The remaining 3 in-distribution and 2 out-of-distribution tasks leave room for improvement. Gemini performs similarly to Med-Gemini-2D across all tasks, although Med-Gemini trends higher in terms of mean AUC(%) on 7 of the 9 in-distribution tasks, and both out-of-distribution tasks.

| Dataset | AugReg-ImageNet | Gemini Ultra | Med-Gemini | PathSSL (Lai et al., 2023) |
|-----------------|--------------------------|--------------------------|--------------------------|----------------------------|
| CAMELYON16 | 95.09 (93.38 - 96.58) | 96.76 (95.13 - 98.13) | 98.48 (97.63 - 99.17) | 99.00 (98.26 - 99.53) |
| Lung AD | 86.04 (82.01 - 89.25) | 87.90 (84.19 - 90.48) | 89.57 (87.30 - 90.93) | 94.41 (91.72 - 96.21) |
| Gleason NCB | 80.79 (78.04 - 83.62) | 84.85 (82.42 - 87.37) | 89.14 (87.29 - 91.04) | 90.59 (88.73 - 92.45) |
| Gleason RP | 85.43 (83.12 - 87.72) | 87.71 (85.30 - 89.93) | 89.39 (87.22 - 91.39) | 91.50 (89.77 - 93.10) |
| Breast IC | 89.47 (88.45 - 90.45) | 91.66 (90.71 - 92.53) | 92.18 (91.24 - 93.07) | 94.32 (93.52 - 95.07) |
| Breast NP | 68.75 (66.66 - 70.73) | 74.68 (72.70 - 76.59) | 73.54 (71.32 - 75.71) | 75.78 (73.50 - 77.90) |
| Breast TF | 74.49 (72.41 - 76.52) | 78.25 (76.25 - 80.15) | 75.66 (73.71 - 77.58) | 83.34 (81.64 - 84.99) |
| CIN | 86.17 (84.49 - 87.77) | 88.22 (86.58 - 89.79) | 89.49 (87.93 - 91.03) | 89.70 (87.96 - 91.30) |
| CRC | 97.29 (96.03 - 98.30) | 98.45 (97.42 - 99.22) | 98.54 (97.70 - 99.22) | 98.95 (98.02 - 99.60) |
| TCGA Study Type | 86.65 (85.53 - 87.75) | 91.85 (91.04 - 92.60) | 92.30 (91.47 - 93.05) | 96.38 (95.83 - 96.88) |
| Tissue Type | 89.37 (88.50 - 90.19) | 92.25 (91.51 - 92.94) | 92.70 (92.00 - 93.34) | 94.76 (94.16 - 95.30) |

A.3. Evaluation metrics

Beyond human and expert evaluation, we leverage a range of automated metrics tailored to specific tasks. For classification tasks, this may include basic accuracy and AUC (Area Under the ROC Curve) metrics. For tasks like report generation, where the fidelity and informativeness of the generated text are crucial, we employ wide variety of metrics such as BLEU, Rouge-L or RadGraph F1-score to probe the quality of our models.

Accuracy Used for image classification and close-ended VQA inference tasks. Measures the percentage of correct predictions *vs.* the ground truth.

AUC (Area Under the ROC Curve) AUC is a performance metric for classification models that indicates how well a model distinguishes between different classes. AUC is calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The TPR measures the proportion of correctly identified positive instances, while the FPR measures the proportion of incorrectly identified negative instances. The area under this curve represents the model's overall ability to separate classes. An AUC of 1.0 indicates a perfect classifier, while an AUC of 0.5 implies the model has no better discriminative power than random guessing.

F1 Score The F1 score is a valuable metric for evaluating classification models, especially when dealing with imbalanced datasets. F1 score is calculated as the harmonic mean of precision (the proportion of true positive out of all predicted positives) and recall (the proportion of true positives correctly identified). The Weighted F1 Score which is used for VQA, averaging F1 scores across classes based on their frequency. Macro-F1 score used for image classification averaging F1 scores across classes without considering imbalances.

Sensitivity Used for image classification in ophthalmology related tasks. Measures the percentage of correctly identified positive cases out of all actual positive cases. A model with high sensitivity minimizes false negatives.

Specificity Used for image classification in ophthalmology related tasks. Measures the percentage of correctly identified negative cases out of all actual negative cases. A model with high specificity minimizes false positives.

Tokenized F1 Tokenized F1-score provides a granular evaluation of language models by calculating precision, recall, and F1-score at the individual token level. This means it rewards partial matches, recognizing the model's ability to identify elements within a sequence even if they're not perfectly aligned. For this purpose True positives and false positives are determined as the number of correctly generated tokens and tokens generated but not present in the ground truth, respectively. False negatives are tokens present in the ground truth but missed by the model.

Rouge-L Rouge-L measures evaluates the quality of generated text and text summarization by comparing the longest common subsequence (LCS) between generated and reference text (Lin, 2004). Higher scores indicate better content and better salient point capturing. Rouge-L assesses the similarity between generated and reference text by measuring the overlap of their LCS and calculating recall based on the LCS length relative to the reference text This metric takes into account the order of words in the text, which makes it particularly suitable for evaluating summaries or text generation tasks where the order of words matters. The higher the Rouge-L score, the better the quality of the generated text compared to the reference text. ROUGE-L relies heavily on LCS and exact matches limiting the contextual understating of the generated text and increasing the sensitivity to sentence length. A high ROUGE-L score doesn't necessarily ensure that the generated text is grammatically correct, well-structured, or reads naturally.

CIDEr CIDEr (Consensus-based Image Description Evaluation) is a metric specifically designed to assess the quality of captions generated for images and short text passages. It goes beyond simple word overlap by considering both the n-gram matches (sequences of consecutive words) and the importance of those n-grams (Vedantam et al., 2015). In the preprocessing, both generated and reference texts are converted to lowercase and common stop words (“the”, “a”, “an”) are removed. Words are also stemmed, reducing them to their root form (e.g., “running” becomes “run”). Then every generated text is broken down into a series of n-grams which are sequences of ‘n’ consecutive words. A weight is assigned to each n-gram based on its Term Frequency-Inverse Document Frequency (TF-IDF). This means common n-grams across all texts receive lower weights, while those that are more informative and distinctive get higher weights. The cosine similarity is calculated between the TF-IDF weighted n-gram vectors of the generated text and each reference. The individual similarity scores are averaged to produce the final CIDEr score. CIDEr can struggle to recognize texts that are semantically similar but use different synonyms and suffer from limited contextual understanding.

BLEU score The BLEU (Bilingual Evaluation Understudy) score is a widely used metric for evaluating the quality of AI generated text. It essentially compares a generated text to a set of human-written reference, providing a score that indicates how similar they are (Papineni et al., 2002). BLEU focuses on n-gram precision, meaning it checks how often sequences of n consecutive words in the generated text appear in any of the reference. It also considers a brevity penalty to discourage generations that are significantly shorter than the reference text. Higher BLEU scores indicate better translation quality, with a perfect score of 1.0 signifying a perfect match between the generated text and the reference. BLEU score has limitations including lack of penalization for grammatical correctness, fluency, or semantic equivalence. Additionally, the quality of the reference and ground truth can impact the BLEU score.

RadGraph F1-score RadGraph F1-score (Jain et al., 2021) is a performance metric specifically designed to evaluate the accuracy of models that extract structured medical information from radiology reports. Unlike standard F1-scores, RadGraph F1-score considers not only whether a finding is correctly identified but also the accuracy of its relationships with other findings within the report. This is crucial because radiology reports often describe complex relationships between abnormalities, locations, and other attributes. Although RadGraph F1-score has its shortcomings, in comparison to other automated NLG metrics provides a more holistic assessment of a model’s ability to understand the nuanced information present in free-text radiology reports.

While RadGraph F1-score offers a more nuanced evaluation than standard F1-scores for radiology report analysis, it has potential limitations. First, it relies on accurate RadGraph creation from the original text. Errors in entity extraction or relation identification during this pre-processing stage could cascade into the RadGraph F1-score calculation. Secondly, it might be overly strict for partial matches and slight discrepancies in relationships or minor variations in wording could significantly penalize the score. Finally, it may not fully account for the clinical relevance of certain errors, treating all mismatches equally despite the potential for varying real-world impact.

To compute the RadGraph F1-score, the model’s predictions on chest X-ray images are compared against ground-truth report made by radiologists or other experts. To increase robustness of our calculation to slight format changes, before passing the ground-truth and the generated report through the RadGraph F1-score package (Yu et al., 2023), we normalize both free-form text to lowercase. The F1-score takes into account both false positives (cases where the model incorrectly identifies an abnormality) and false negatives (cases where the model fails to detect a true abnormality). By considering both precision (the ratio of true positives to the total number of predicted positives) and recall (the ratio of true positives to the total number of actual positives), the F1-score provides a balanced assessment of the model’s performance. A higher RadGraph F1-score indicates better performance in accurately identifying abnormalities in medical images, which is crucial for assisting radiologists in diagnosis and treatment planning.

A.4. Supplementary Table for Performance Summary

Table A.15 presents the aggregate performance of Med-Gemini compared to the previous state-of-the-art (SoTA), or a strong baseline where available. Figure 1 illustrates the relative improvement gained by using one of our Med-Gemini models over the SoTA or strong baseline, using Gemini as a reference point when no SoTA is available. For pathology classification, we averaged AUC performance across all sub-datasets. For report generation, we calculated the micro average performance across normal and abnormal cases, expert identified “AI generated report is superior or similar to original report” (see Table 7)

Table A.15 | Overall Performance Summary of Med-Gemini This table represent the aggregated results comparing Med-Gemini to the previous state-of-the-art (SoTA), Gemini or strong baseline where available.

| Capabilities | Datasets | Metric | Med-Gemini | Gemini | Baseline or SoTA | Reference |
|----------------------------|-------------------------|----------------------|------------|--------|------------------|---|
| Report Generation | MIMIC-CXR | RadGraph | 24.4 | N/A | 20.5 | Tanno et al. (2024) |
| | MIMIC-CXR | Expert (AI superior) | 47.6 | N/A | 43.0 | Tanno et al. (2024) |
| | IND1 | Expert (AI superior) | 75.4 | N/A | 63.7 | Tanno et al. (2024) |
| VQA | MIMIC-CXR VQA | Accuracy | 78.6 | 70.9 | 68.1 | Xu et al. (2023) |
| | Slake-VQA | Accuracy | 84.8 | 70.4 | 91.1 | Li et al. (2023b) |
| | VQA-Rad CXR | Expert | 71.9 | N/A | 57.9 | Xu et al. (2023) |
| | VQA-Rad CXR | Accuracy | 78.8 | 62.4 | N/A | N/A |
| | PathVQA | Accuracy | 83.3 | 62.8 | 90.9 | Sun et al. (2024) |
| Genomic Disease Prediction | Coronary artery disease | AUC | 82.5 | N/A | 78.5 | Ensemble of PRSs and Demographics |
| | Stroke | AUC | 78.1 | N/A | 74.8 | |
| | Type 2 diabetes | AUC | 86.5 | N/A | 85.5 | |
| | Glaucoma | AUC | 78.9 | N/A | 82.8 | |
| | COPD | AUC | 83.9 | N/A | 77.4 | |
| | Rheumatoid arthritis | AUC | 79.2 | N/A | 70.6 | |
| | Major depression | AUC | 70.2 | N/A | 60.2 | |
| | Allcause mortality | AUC | 79.9 | N/A | 71.8 | |
| | Hypertension | AUC | 72.0 | N/A | 69.9 | |
| | Hypercholesterolemia | AUC | 52.4 | N/A | 52.1 | |
| | Atrial fibrillation | AUC | 78.7 | N/A | 74.4 | |
| Diabetic retinopathy | AUC | 87.0 | N/A | 87.5 | | |
| | Pneumonia | AUC | 71.6 | N/A | 68.9 | |
| | Asthma | AUC | 57.7 | N/A | 56.5 | |
| Classification | Hard Exudates | F1 | 87.3 | 61.5 | N/A | N/A |
| | Hemorrhage | F1 | 85.3 | 57.8 | N/A | N/A |
| | PRP Scars | F1 | 82.3 | 57.0 | N/A | N/A |
| | DR Lesion | F1 | 86.4 | 63.9 | 92.0 | Krause et al. (2018) |
| | PAD-UFES-20 | F1 | 71.4 | 60.3 | 70.0 | Google (2024) |
| | MIMIC-CXR | F1 | 90.7 | 86.8 | N/A | N/A |
| | CheXpert | F1 | 48.3 | 42.6 | 60.6 | Tiu et al. (2022) |
| | ChestX-ray14 | F1 | 46.7 | 34.2 | 58.3 | Majkowska et al. (2020) |
| | Pathology Patch | AUC | 89.2 | 88.4 | 91.7 | Sun et al. (2024) |