

MedGemma Technical Report

Google Research and Google DeepMind ¹

Artificial intelligence (AI) has significant potential in healthcare applications, but its training and deployment are challenging due to healthcare's diverse data, complex spectrum of possible tasks, and the need to preserve privacy. Foundation models that perform well on various medical tasks and require less task-specific tuning data are critical to accelerating the development of AI for healthcare applications. In this technical report, we introduce MedGemma, a new collection of medical vision-language foundation models based on Gemma 3 4B and 27B. MedGemma demonstrates advanced medical understanding and reasoning on images and text, significantly exceeding the performance of similar-sized generative models and approaching the performance of task-specific models, while maintaining the general capabilities of the Gemma 3 base models. For out-of-distribution tasks, MedGemma achieves 2.6-10% improvements on medical multimodal question answering, 15.5-18.1% improvements on chest X-ray finding classification, and 10.8% improvement on agentic evaluations compared to the base models. Fine-tuning MedGemma further improves performance in subdomains, reducing errors in electronic health record information retrieval by 50% and reaching comparable performance to existing specialized state-of-the-art methods for pneumothorax classification and histopathology patch type classification. We additionally introduce MedSigLIP, a medically-tuned vision encoder derived from SigLIP. MedSigLIP powers the visual understanding capabilities of MedGemma and, as an encoder, it achieves performance comparable to or better than specialized medical image encoders. Taken together, the MedGemma collection provides a strong foundation of medical image and text capabilities, with potential to significantly accelerate medical research and development of downstream applications. More details about the MedGemma collection, including tutorials and instructions for downloading the model weights, can be found at <https://goo.gle/medgemma>.

¹ See Contributions and Acknowledgments section for full author list.
Corresponding authors: {linyan, dangolden, asellerg}@google.com.

1. Introduction

The landscape of modern healthcare is characterized by the generation and use of an unprecedented volume and diversity of data. Diagnosis, treatment, and monitoring rely on synthesizing information from disparate sources and specialties. Recently developed large multimodal models (LMMs), trained on massive and diverse datasets, exhibit remarkable capabilities in detecting complex patterns, generating coherent text, and processing visual information (Achiam et al., 2023; Alayrac et al., 2022; Chen et al., 2022; Liu et al., 2023, 2024; OpenAI, 2023; Touvron et al., 2023). These capabilities mark a potential paradigm shift in assisting with current workflows and extracting novel insights.

While general-purpose (non-medically tuned) LMMs demonstrate impressively broad abilities, generic models can lack nuanced medical understanding and the ability to interpret and reason about medical data in a robust way (Han et al., 2023; Labrak et al., 2024; Singhal et al., 2023b,c; Toma et al., 2023; Tu et al., 2024; Yang et al., 2024). Recognizing this gap, we created MedGemma, a new suite of open, medically-tuned, vision-language foundation models. These models represent the latest addition to the Health AI Developer Foundations (Kiraly et al., 2024) collection. Built upon the robust architecture of Gemma 3 (Gemma-Team et al., 2025), the MedGemma models are designed to interpret and reason about medical images and text while retaining the strong general-purpose capabilities present in Gemma 3.

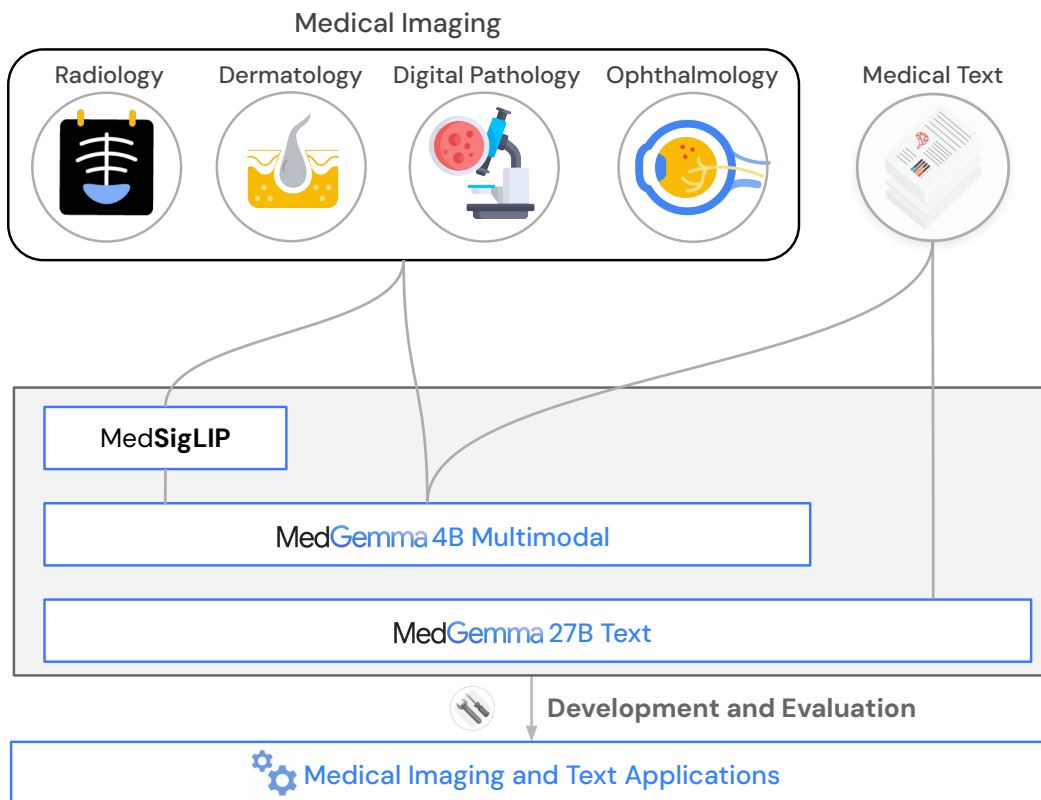


Figure 1 | Overview of the MedGemma model collection featuring the MedSigLIP image encoder, MedGemma 4B Multimodal and MedGemma 27B Text

In this report, we focus on two MedGemma models: a 4B variant that can accept text, images, or both as input, and a 27B variant that is optimized for text-only inputs. Both models output text. MedGemma 4B demonstrates strong performance on Vision Question Answering (VQA) benchmarks compared to prior SOTA models like Med-Gemini (Saab et al., 2024; Yang et al., 2024) despite being considerably smaller. Both MedGemma 4B and 27B are highly competitive on challenging

text-only medical benchmark tasks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU Med (Hendrycks et al., 2020), AfriMed-QA (Olatunji et al., 2024), and AgentClinic (Schmidgall et al., 2024) when compared against other open models of similar scale. In addition to these strong out-of-the-box capabilities, we show how performance can be further improved by fine-tuning MedGemma on subdomains like chest X-ray reporting, histopathology classification, and electronic health record information retrieval.

An additional MedGemma variant, a multimodal version of MedGemma 27B, was also developed and is being released along with the other models. More thorough evaluation of this multimodal 27B variant is ongoing and preliminary results can be found in Appendix Section F. Unless otherwise noted in this report, evaluations that reference “MedGemma 27B” refer to the text-only variant of MedGemma 27B.

In addition to the MedGemma models, we describe the standalone MedSigLIP 400M-parameter medical image encoder. MedSigLIP is based on SigLIP-400M (Zhai et al., 2023) and is the same encoder that powers MedGemma’s image interpretation capabilities. When used on its own, MedSigLIP enables data-efficient and zero-shot image classification and retrieval, with performance comparable to or exceeding specialized image encoders.

A high level overview of the released models is shown in Fig. 1. More details about the MedGemma collection, including tutorials and links to download all of the above models, can be found at <https://goo.gle/medgemma>.

2. Methods

2.1. Datasets

For general purpose data replay during pretraining, original data mixtures from SigLIP (Zhai et al., 2023) and Gemma 3 (Gemma-Team et al., 2025) were leveraged. The medical training and evaluation datasets largely followed the datasets in Med-Gemini (Yang et al., 2024). In this section, we outline the specific changes or differences in datasets relative to Med-Gemini.

2.1.1. Training datasets

Text-only datasets: For text datasets, we sampled responses and logits from a large IT (instruction-tuned) teacher using the train splits of multiple medical QA datasets, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MedExpQA (Alonso et al., 2024), AfriMed-QA (Olatunji et al., 2024), HealthSearchQA (Singhal et al., 2023a), and LiveQA (Abacha et al., 2017). We also sampled responses and logits for approximately 200,000 synthetic medical questions generated by asking the same large IT teacher to generate a new question using 5 randomly sampled questions from the above datasets as examples.

Multimodal datasets: Relative to Med-Gemini, the multimodal capabilities of MedGemma are currently focused on 2D medical images (e.g. X-ray, 2D slices from CT/MRI); 3D volumes and genomic datasets described in Yang et al. (2024) were not included. Additionally, we and others have identified potential data quality issues in PathVQA and MedVQA. Thus, we removed them from the training dataset. We did not include PAD-UFES-20 in the post-training dataset since it focuses on 6-class classification of very specific lesion types, which is not in line with the goal of more general purpose dermatology capabilities and use cases. For the PMC-OA component of the training data, we only included the single panel medical images from PMC-OA for better data quality. Relative to Med-Gemini we also introduced a larger internal collection for ophthalmology (184,852 more retinal fundus images), dermatology (51,049 more dermatology images with 210 different skin conditions), histopathology (a total of ~32.5 million patch-text pairs), and radiology data (54,573 more CT 2D

Table 1 | Overview of MedGemma training datasets.

Modality	Dataset	No. examples	Training stages	Description
Text-only	MedQA	9,275	Distill, RL	USMLE style exam questions
	MedMCQA	182,806	Distill, RL	Indian medical entrance exam questions
	AfriMed-QA	1,003	Distill, RL	Pan-African English multi-specialty QA
	MedExpQA	434	Distill, RL	Spanish medical residency exam questions
	PubMedQA	1,000	Distill, RL	Biomedical QA compiled from PubMed abstracts
	LiveQA	634	Distill, RL	Consumer health questions from United States National Library of Medicine (NLM)
	HealthSearchQA	3,375	Distill, RL	Common consumer medical questions from search engines
	Synthetic	200,000	Distill	Generated from large IT teacher
Radiology	SLAKE	450	Vision, PT, RL	Captions generated from QA pairs
	VQA-Rad	1,391	RL	Radiology image QA pairs
	MIMIC-CXR	231,483	Vision, PT, RL	Chest X-ray images & free-form reports
	Digital Knee X-ray	1,469	Vision, PT	Knee X-ray images & labels
	CT-US1	59,979	Vision, PT	2D CT slices & free-form reports
	MRI-US1	47,622	Vision, PT	2D MRI slices & free-form reports
Histopathology	Internal histopathology	32,550,599	Vision, PT, RL	Histopathology image patches, caption pairs
Dermatology	PAD-UFES-20	2,047	Vision, PT	Skin lesion images & labels
	Internal dermatology	51,049	Vision, PT, RL	Skin lesion images & labels
Ophthalmology	EyePACS	199,258	Vision, PT, RL	Fundus images & labels
General Medical	PMC	41,853	Vision, PT	Single panel medical images & caption pairs

Vision: Vision encoder enhancement, PT: Pretraining, Distill: Distillation, RL: Reinforcement learning.

slices, 47,622 more MRI 2D slices). The additional CT and MRI slices utilized for training were curated based on mention of a specific slice associated with abnormal findings in the radiology report.

2.1.2. Data Preprocessing

Our data preparation followed [Yang et al. \(2024\)](#) closely. Image padding and resizing algorithms remain the same, but because the vision encoder is different in Gemma 3, our images were resized to 896×896 instead of 768×768. Following Gemma 3, we use the SentencePiece tokenizer with 262,000 entries. Additionally, for CT images, we preselected three windows and converted them into the RGB color channels of the input image to highlight (1) bone and lung, window-width: 2250, window-level: -100; (2) soft tissue, window-width: 350, window-level: 40; (3) brain, window-width: 80, window-level: 40.

2.2. Modeling Methodology

2.2.1. Modeling Architecture and Training Infrastructure

The MedGemma model architecture follows Gemma 3 ([Gemma-Team et al., 2025](#)) and is compatible with all existing Gemma infrastructure. The vision encoder for Gemma 3 is the 400M variant of the SigLIP encoder ([Zhai et al., 2023](#)) and is shared across the different Gemma language model sizes (4B, 27B). The input image resolution is 896×896 with pixel values normalized to [-1, 1]. The language model component also follows Gemma 3, featuring arbitrary image-text interleaving and long context (128k). Similar to Gemma 3, MedGemma was trained on TPUv4, TPUv5e, and TPUv5p, leveraged pre-computed visual tokens for memory saving, and used data and model shardings for multi-pod training.

2.2.2. Model Training

The MedGemma 4B multimodal model utilized all of the following steps while the text-only version of MedGemma 27B leveraged the post-training stage alone.

Vision Encoder Enhancement for MedGemma: To improve the vision encoder’s capability of encoding and distinguishing subtle differences in medical images, we fine-tuned the vision encoder in Gemma 3 (SigLiP-400M) using over 33M medical image-text pairs (635k from various medical modalities and 32.6M histopathology patches) as listed in Table 1. To retain SigLiP’s existing performance, its original training data (e.g., WebLI) were retained and medical data was mixed with 2% weight into the training. While the Gemma 3 vision encoder works with 896×896 resolution, we found that many medical vision tasks worked reasonably well at 448×448 resolution (Table 15). Thus, while the MedGemma 4B image encoder is based on 896×896 resolution for compatibility and consistency with Gemma 3, the released MedSigLiP model is based on 448×448 resolution for more efficient experimentation and adaptation by the community. The 448×448 encoder shares the same model weights as the 896×896 encoder with the only difference being down-sampled positional embeddings to work with fewer input patches from the lower resolution.

Multimodal Decoder Pretraining: After the vision encoder enhancement, the Gemma language model needed to be re-adapted for this new vision encoder, not only for the medical data but also for the general image domain to preserve the visual-language reasoning capabilities. This goal was achieved in the pretraining stage incorporating both the text and interleaved imaging data from the original mixture and the newly introduced medical domain image-text paired data. Notably, we did not introduce further medical text-only data in this step, as the original Gemma 3 mixture was already general-purpose and large-scale. To reduce compute requirements, we continued our pretraining on top of the original Gemma 3 pretrained checkpoints, mixed our medical image data (Table 1) with 10% weight, trained for approximately 5 epochs on the medical mixture given the mixing ratio, and picked the checkpoint based on the validation set performance on chest X-ray report generation, and radiology, dermatology, and ophthalmology visual question answering.

Post-training: The knowledge acquired from pretraining needs to be surfaced as capabilities in the post-training stage. There are two primary post-training components as previously outlined for Gemma 3. The recipes for distillation and reinforcement learning (RL) were the same as in Gemma 3 development with the following additions: (1) Distillation: addition of medical text data during this component to enable further learning in these domains from a large instruction-tuned (IT) teacher. (2) Reinforcement learning: Medical imaging data with paired text was utilized in the RL stage of post-training. For multimodal training, we found that RL enables better generalization compared to supervised fine-tuning, so all multimodal post-training was performed via RL.

3. MedGemma Evaluations

MedGemma was evaluated and compared with other models on five types of medical tasks: text question-answering, image classification, visual question answering, chest X-ray (CXR) report generation, and agentic behavior. We additionally validated MedGemma on several general purpose (non-medical) benchmarks. A high-level overview of tasks and datasets is provided in Table 2 and additional details on each task are below.

3.1. General evaluation approach

Evaluation parameters: Unless reported otherwise, all evaluations that we performed consisted of a single inference run per example. For MedGemma evaluations, a temperature of 0.0 was used on medical benchmarks and the default temperature was used on non-medical benchmarks. For

Table 2 | Overview of MedGemma evaluation datasets.

Task	Dataset	Modality	No. Examples	OOD [†]
Medical text question-answering	MedQA	Text	1,273	-
	MedMCQA	Text	4,183	-
	PubMed QA	Text	500	-
	MMLU Med	Text	3,685	-
	MedXpertQA	Text	2,450	✓
	AfriMed-QA (MCQ)	Text	25	-
Medical image classification	MIMIC-CXR (Med-Gemini test set)	Radiology	1532	-
	MIMIC-CXR (MAIRA test set)	Radiology	2461	-
	ChestX-ray14 (CXR14)	Radiology	1,962	✓
	CheXpert	Radiology	668	✓
	US-Derm MCQA	Dermatology	1,996	-
	Path MCQA	Histopathology	450	-
	EyePACS	Ophthalmology	3,161	-
Medical visual question-answering	MedXpertQA	General medical	2,000	✓
	SLAKE (English-only)	Radiology	1,061	-
	VQA-RAD	Radiology	2,248	-
Chest X-ray report generation	MIMIC-CXR	Radiology	306	-
Medical agentic behavior	AgentClinic-MedQA	Agentic text	215	-
	AgentClinic-MIMIC-IV	Agentic text	200	✓
General purpose	MMLU Pro	Text	12,032	N/A
	Global MMLU Lite	Text	6,400	N/A
	MMMU (val)	Text + Image	900	N/A

[†] Out of Distribution: Data not seen during any model development stages. For general purpose benchmarks, it is difficult to determine if data are OOD given the large amount of pretraining data in the original mixture.

evaluations of all other models, on all datasets, each model’s default temperature and top-k were used. Due to data privacy and license terms, only publicly available datasets were used in evaluating models involving public APIs (e.g. OpenAI models). In cases where existing literature with performance metrics was available, those values were used and noted (with inclusion criteria for external models described below). For generalist models, we found that giving them a persona in the system message, such as “You are a helpful medical assistant” or “You are a helpful radiology assistant” could improve their performance, we thus added these messages into the evaluation prompts. Detailed prompt usages can be found in Appendix Tables A6 and A7. As the DeepSeek R1 model (DeepSeek-AI, 2025) is a text-only model, it was only evaluated on text benchmarks.

Inclusion of previously published model performance data: Where comparisons to other models were made, we restricted inclusion to models that met the following criteria: a publicly accessible model card (e.g., via Hugging Face or an institutional website); clear and verifiable attribution, including the name and contact information of the responsible individual or institution; explicit licensing terms governing its use; at least one associated technical report or publication for the current or a prior version of the model. These inclusion criteria were chosen to help ensure accountability, transparency, and adherence to sound machine learning practices, such as avoiding test dataset leakage. Additionally, for visual question answering (VQA) comparisons, we only include zero-shot generative evaluations for the most direct and meaningful comparisons (thus excluding few-shot or discriminative, embedding-based approaches for VQA).

3.2. Medical text question-answering

For evaluation of medical and health related capabilities, we used the official, publicly available test splits for MedQA, MedMCQA, PubMedQA, MMLU medical subcategories, AfriMed-QA, and

MedXpertQA. AfriMed-QA includes a mix of closed and open questions though we limited evaluations to only the closed multiple choice questions (MCQs). No data from MedXpertQA was used in model training, so it is considered an out-of-distribution (OOD) benchmark.

3.3. Medical image classification

We evaluated medical image classification on three public chest X-ray datasets, as well as private datasets of dermatology, histopathology and retinal fundus images. We measured classification performance using accuracy or macro F1, depending on the dataset, targeting common findings or diagnoses in each modality.

Chest X-rays: Prediction accuracy was evaluated for five conditions in the MIMIC-CXR (Goldberger et al., 2000; Johnson et al., 2019a,c) and CheXpert (Irvin et al., 2019) datasets: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. Accuracy was evaluated for three conditions in the ChestX-ray14 (CXR14) (Wang et al., 2017) dataset: lung opacity, pneumothorax, and fracture. We used the original condition labels from CheXpert. For MIMIC-CXR, we report performance on two different versions of the test data: (1) using radiologist-adjudicated labels with missing and uncertain labels excluded, as in Yang et al. (2024) and (2) using original labels from Johnson et al. (2019b) with missing and uncertain labels considered to be negative, with the same set of test cases as reported in Hyland et al. (2023). For the CXR14 data set, radiologist-adjudicated labels as described in Majkowska et al. (2020) were used for evaluation. Evaluations on MIMIC-CXR were not performed with the OpenAI o3 model due to data privacy considerations.

Dermatology: “US-Derm MCQA” (Liu et al., 2020) is an internal, de-identified dataset consisting of one image per patient from 1996 patients who were referred to tele-dermatologists by primary care physicians in the United States. There are 136 different skin conditions represented across the images, with ground truth diagnoses provided by dermatologists based on the images and metadata. We converted this dataset into a multiple choice question format where the associated reference condition is included among three other randomly assigned condition labels from the same dataset (four options total per MCQ).

Histopathology: “Path MCQA” is an internal dataset of 450 patches extracted from 354 unique whole slide images associated with specimens from breast cancer, lung cancer, prostate cancer, lymph nodes, and cervical biopsies. These represent test splits (by patient) from several different data sources. Patches comprise magnifications of 5x, 10x, and 20x (2, 1, and 0.5 microns per pixel). A single, multiclass labeling task for each patch was formulated as a multiple choice question, with four to nine possible options per question depending on the tissue type and labeling task associated with the image. These questions focus on identification and grading for breast cancer, prostate cancer, and cervical dysplasia as well as lung cancer histologic classification and histologic sub-typing. Ground truth labels were obtained via annotations provided as region-level labels by US board-certified pathologists (Jaroensri et al., 2022; Nagpal et al., 2019, 2020; Sadhwani et al., 2021).

Ophthalmology: We used the de-identified 45-degree fundus imagery dataset from EyePACS (Cuadros and Bresnick, 2009) as described previously (Yang et al., 2024). We evaluated one image per patient from 3161 patients on clinically-determined 5-class diabetic retinopathy (DR) severity labels, with each task formulated as a multiple choice question with five options: none, mild, moderate, severe or proliferative DR.

3.4. Medical visual question-answering

We measured visual question-answering performance on the radiology SLAKE and VQA-RAD data sets using the average tokenized F1 metric across open and closed QAs, as well as accuracy on the subset

of yes/no questions. For SLAKE, we used the default train/test split, and for VQA-RAD, we used splits from Yang et al. (2024)² to avoid the train/test image contamination present in the original splits.

3.5. Chest X-ray report generation

The MedGemma 4B pretrained model was used to generate radiology reports from the MIMIC-CXR test set. The pretrained model was chosen here rather than the post-trained model due to the sensitivity to reporting style of metrics like RadGraph F1 (Jain et al., 2021). The pretrained model could better follow the style of MIMIC-CXR, as MIMIC-CXR reports are used in training, while the post-trained model conformed more closely to the original Gemma 3 style in terms of report generation.

We measured accuracy of chest X-ray report generation on the MIMIC-CXR dataset by comparing MedGemma-generated reports to the original radiologist reports for both impression and findings using the RadGraph F1 metric (Jain et al., 2021) on the 912 image set used in Tanno et al. (2024) and Yang et al. (2024).

We also performed a human expert evaluation on the same 306-case image set as in Yang et al. (2024) with a US board-certified cardiothoracic radiologist to evaluate both the original and MedGemma-generated reports with respect to the associated chest X-ray image. The evaluation task compared the reports on a five-point scale, as originally described in Yang et al. (2024) and shown in Appendix Table A1. Although the reviewer was asked to remain neutral in their evaluation, they were not blinded to which report was from the original radiologist vs. from the AI system. This evaluation complements the automated RadGraph-based evaluation as it can both distinguish between major and minor issues as well as account for scenarios in which the original MIMIC-CXR report contains errors or omissions.

3.6. Medical agentic behavior

To provide insights into the capabilities of MedGemma in more complex environments, we measured the ability of MedGemma to operate in an agentic setting. We evaluated on AgentClinic (Schmidgall et al., 2024), which positions MedGemma in the role of a “physician agent” in a simulated clinical environment. Solving tasks in AgentClinic required the model to perform dialogue-driven patient history taking, ordering and interpreting medical exams, and operating under incomplete information in order to accurately provide a final diagnosis. We evaluated on 415 simulated cases using the text-only environments, AgentClinic-MedQA (215 cases) and AgentClinic-MIMIC-IV (200 cases), which are derived from MedQA (Jin et al., 2021) and MIMIC-IV (Johnson et al., 2023) respectively.

3.7. General purpose benchmarks

Given the limitations that many specialized medical models exhibit when faced with non-medical tasks, we also evaluated possible tradeoffs of specialization by evaluating on the MMLU Pro, Global MMLU Lite, and MMMU benchmarks. Performance on the MMMU benchmark is reported on the validation set as the public test set does not include answers.

4. MedGemma Results

Medical text question-answering: Across all text-only biomedical QA tasks evaluated, MedGemma demonstrated superior performance over the standard Gemma 3 model variant of the same size, as well as competitive performance with much larger models in many cases. This was based on evaluation on MedQA, MedMCQA, PubMedQA, MMLU subsets and AfriMed, shown in Table 3, and the out-of-distribution MedXpertQA, shown in Table 4.

²https://github.com/Google-Health/google-health/blob/master/data_splits/

Table 3 | Accuracy on text-only medical benchmarks. Metrics for MedQA correspond to the original 4 option test set, unless otherwise specified with an asterisk. Metrics for external, small models as well as OpenBioLLM 70B were obtained from the prior reports for these models. Other metrics were computed internally as described in the evaluation section.

Model [†]	Open [‡]	MedQA	MedMCQA	PubMedQA	MMLU Anatomy	MMLU Clinical Knowl.	MMLU College Biology	MMLU College Medicine	MMLU Medical Genetics	MMLU Prof. Medicine	MMLU Virology	AfriMed
Small Models												
MedGemma 4B	✓	64.4	55.7	73.4	59.3	71.3	70.8	65.3	83.0	76.8	53.0	52.0
Gemma 3 4B	✓	50.7	45.4	68.4	54.1	69.8	77.8	63.0	74.0	65.4	42.8	48.0
MedGemma 27B (with test-time scaling)	✓	87.7	74.2	76.8	83.7	86.0	96.5	86.1	97.0	93.4	53.6	84.0
Gemma 3 27B	✓	74.9	62.6	73.4	74.8	86.0	93.8	78.6	91.0	85.7	51.2	72.0
BioMistral DARE 7B [§] (Labrak et al., 2024)	✓	51.1	48.7	77.7	55.8	62.3	66.9	58.0	67.0	61.4	N/A	N/A
JSL-MedLlama 3 8B v2.0 [§] (John Snow Labs, 2024)	✓	61.4*	61.2	74.2	71.9	78.1	82.6	71.1	83.0	78.7	N/A	N/A
OpenBioLLM 8B [§] (Ankit Pal, 2024)	✓	59.0	56.9	74.1	69.8	76.1	84.2	68.0	86.1	78.2	N/A	N/A
IQVIA Med-R1 8B [§] (IQVIA, 2025)	-	73.3	63.3	76.4	72.6	78.5	88.2	72.8	87.0	84.9	N/A	N/A
Large Models												
OpenBioLLM 70B [§] (Ankit Pal, 2024)	✓	78.2	74.0	79.0	83.9	92.9	93.8	85.7	93.2	93.8	N/A	N/A
DeepSeek R1 (DeepSeek-AI, 2025)	✓	90.1	78.8	77.2	91.1	91.7	98.6	90.8	99.0	95.6	56.0	92.0
Gemini 2.5 Flash	-	92.0	79.7	76.2	91.1	91.7	98.6	87.9	97.0	96.0	59.6	84.0
Gemini 2.5 Pro	-	92.6	81.1	75.8	91.1	91.7	98.6	89.0	96.0	96.3	56.0	84.0
GPT-4o	-	86.5	76.1	78.4	86.7	89.4	94.4	86.7	98.0	93.0	58.4	80.0
o3	-	93.3	83.3	80.0	91.9	94.7	98.6	90.2	100	96.0	56.0	84.0

* Indicates results for the all-options version of the dataset

[†] See Section 3.1 for inclusion criteria

[‡] Open-weight models

[§] Prior reported results

|| Research use only

Medical image classification: To further evaluate MedGemma across the modalities that are most highly represented in the training data, we utilized a set of image-based classification tasks across radiology, histopathology, dermatology, and ophthalmology. Although addressing classification as a zero-shot generative task may not provide maximum performance compared to training an embedding-based classifier, these evaluations provide additional insights into the quality and performance of underlying image encoder and the overall model capabilities. Results are summarized in Table 7 for CXR evaluations and Table 8 for histopathology, dermatology, and retina image classification. MedGemma demonstrated substantially superior performance on these tasks relative to the Gemma 3 baselines, and notably, MedGemma also demonstrated superior performance as compared to much larger API-based models.

Table 4 | Accuracy results on MedXpertQA (OOD)

	Small Models				Large Models		
Type	MedGemma 4B	Gemma 3 4B	MedGemma 27B	Gemma 3 27B	Gemini 2.5 Flash	Gemini 2.5 Pro	o3
Text-only	14.2	11.6	25.7	15.7	36.2	43.1	54.6
Multi-modal only	24.4	22.3	N/A	29.8	47.4	58.9	67.5

MedGemma is also capable of engaging in open-ended clinical reasoning tasks. Examples of such tasks with MedGemma 27B, including clinical commentary on model performance, are shown in Table 5 (long-form MedGemma response) and Table 6 (succinct MedGemma response). Additional examples can be found in Appendix Section E.

Medical visual question answering: Performance for SLAKE (Liu et al., 2021) and VQA-RAD (Lau et al., 2018) are summarized in Table 9. Again, MedGemma 4B demonstrated superior performance over the standard Gemma 3 model variant of the same size, as well as competitive performance with much larger models in many cases. Of note, we did not compare to models for which these metrics are calculated using a discriminative, embedding-based approach. Such approaches can be informative, but use the train set to define the set of possible text answers, and thus represent a classification

Table 5 | Example of MedGemma 27B clinical reasoning in an oncology setting, including commentary by a US board-certified practicing oncologist

Prompt

A 62-year old man presents with a 2 cm spiculated right upper lobe peripheral lung nodule and mediastinal lymphadenopathy. Imaging furthermore shows multiple metastatic lesions in the liver and bones, and a solitary 1 cm lesion in the right parietal lobe without surrounding vasodema. A biopsy from the liver demonstrated adenocarcinoma consistent with a lung primary. Genomic sequencing demonstrated TP53 and EGFR L858R mutation, and PD-L1 was 70%. On examination, the patient is without any neurologic deficits and is ECOG 0. What is the best next step in management for this patient?

MedGemma 27B output

Okay, let's break down the management for this patient with Stage IV NSCLC.

Key Findings:

1. **Diagnosis:** Stage IV Adenocarcinoma of the Lung (confirmed by liver biopsy).
2. **Driver Mutation:** EGFR L858R mutation (highly actionable).
3. **Biomarker:** PD-L1 70% (high expression).
4. **Metastases:** Widespread (mediastinal nodes, liver, bone, solitary brain lesion).
5. **Clinical Status:** Excellent performance status (ECOG 0), asymptomatic brain lesion.

Analysis & Rationale:

- **EGFR Mutation is Paramount:** The presence of the EGFR L858R mutation is the most critical factor guiding first-line therapy. Patients with EGFR-mutated NSCLC have demonstrated significantly improved outcomes with EGFR tyrosine kinase inhibitors (TKIs) compared to chemotherapy.
- **PD-L1 High:** While PD-L1 is high (70%), indicating potential benefit from immunotherapy, current guidelines and clinical trial data (like FLAURA) strongly favor EGFR TKIs as the *first-line* treatment for patients with EGFR mutations, even those with high PD-L1. Immunotherapy is typically considered after TKI failure or in combination with chemotherapy in specific settings, but not usually as the initial monotherapy in this scenario.
- **Brain Metastasis:** The solitary, asymptomatic brain lesion needs management. EGFR TKIs, particularly third-generation agents like Osimertinib, have good central nervous system (CNS) penetration and activity. Therefore, systemic therapy with an EGFR TKI is often the initial approach to treat the brain metastasis as well. Close monitoring with MRI is essential. If the lesion progresses despite systemic therapy or becomes symptomatic, Stereotactic Radiosurgery (SRS) would be considered.
- **ECOG 0:** The patient's excellent performance status makes him a suitable candidate for aggressive systemic therapy.

Best Next Step in Management: The best next step is to initiate **first-line systemic therapy with an EGFR tyrosine kinase inhibitor (TKI)**.

- **Preferred Agent: Osimertinib** is generally the preferred first-line EGFR TKI due to its superior efficacy (progression-free survival and overall survival) and improved CNS penetration compared to earlier-generation TKIs (like Gefitinib, Erlotinib, Afatinib) demonstrated in the FLAURA trial.
 - **Management of Brain Metastasis:** Start Osimertinib. Monitor the brain lesion closely with serial MRI scans (e.g., every 3-6 months initially). Consider SRS if the lesion progresses or causes symptoms.
 - **Supportive Care:** Address potential side effects of the TKI (e.g., rash, diarrhea). Manage potential bone pain with analgesics or palliative radiation if needed.
-

Commentary by US board-certified practicing oncologist

Strengths: Identifies the key components of staging and the actionable driver mutation. PD-L1 is high but correctly identifies TKI as preferable over immunotherapy. Correctly identifies osimertinib as having good brain penetration especially with asymptomatic and small brain lesions. Appropriately highlights FLAURA trial. Includes supportive care aspects including potential side effects.

Areas for Improvement: Immunotherapy can be considered after TKI progression, though the treatment paradigm has evolved (i.e., amivantamab + chemotherapy, datopotamab deruxtecan). Although osimertinib is a correct answer, newer combinations are available and have shown benefit over osimertinib alone (i.e., amivantamab + lazertinib, osimertinib + chemotherapy). Serial brain MRI is mentioned but serial rest-of-body imaging is not. Dosing of osimertinib not mentioned. Bone disease discussed, though no consideration for bone modifying agents or for whether surgery may be indicated. Shared decision-making would be a beneficial addition.

Table 6 | Example of MedGemma 27B clinical reasoning in an oncology setting, specifically prompted for *succinct* output, including commentary by a US board-certified practicing oncologist

Prompt

A 62-year old man presents with a 2 cm spiculated right upper lobe peripheral lung nodule and mediastinal lymphadenopathy. Imaging furthermore shows multiple metastatic lesions in the liver and bones, and a solitary 1 cm lesion in the right parietal lobe without surrounding vasodema. A biopsy from the liver demonstrated adenocarcinoma consistent with a lung primary. Genomic sequencing demonstrated TP53 and EGFR L858R mutation, and PD-L1 was 70%. On examination, the patient is without any neurologic deficits and is ECOG 0. What is the best next step in management for this patient? *Please be succinct.*

MedGemma 27B output

Initiate Osimertinib and monitor the brain lesion with serial MRI.

Commentary by US board-certified practicing oncologist

Strengths: To the point, directly addresses systemic therapy + serial brain MRI.

Areas for Improvement: A little too succinct, does not discuss radiosurgery, and otherwise the same issues as the lengthy output in the prior example.

task instead of a more flexible open-ended VQA task. Additionally, LLaVA-Med and BiomedGPT-B are included in the results table using previously reported metrics (Li et al., 2023; Zhang et al., 2023) and demonstrated very strong performance, with the qualifier that such performance involves supervised fine-tuning for the specific datasets. Additionally, LLaVA-Tri (Xie et al., 2024) has reported even higher performance metrics, but the reported methodological details for these metrics are insufficient to inform direct comparison. In addition to these statistical results, we show several anecdotal examples of open-ended visual question-answering in Figures 2 and 3.

Chest X-ray report generation: Results of performance on CXR report generation based on the MIMIC-CXR dataset and the RadGraph F1 metric are shown in Table 10. The results demonstrate that our pretrained MedGemma models can achieve SOTA-level results by training with MIMIC-CXR dataset as our only CXR-related dataset. This shows the versatility of Gemma’s architecture in adapting to complex medical use cases. In addition, expert human evaluation shown in Figure 4 found that 68% and 49% of generated reports in normal and abnormal studies respectively were equal or superior to the original reports, with 81% of generated reports overall resulting in the same or superior clinical decisions in comparison to the original reports. For comparison, in a similar study, 73% of reports generated by the larger Med-Gemini model would have resulted in the same or superior clinical decisions in comparison to the original reports (Yang et al., 2024).

Medical agentic behavior: To provide initial insights into the capabilities of MedGemma in the setting of more realistic use cases beyond single turn question answering, we also evaluated MedGemma in an agentic role using the AgentClinic benchmark (Schmidgall et al., 2024). In this task, the model functions as a “physician agent” in simulated clinical encounters, requiring tasks like patient history taking, exam interpretation, and diagnosis under uncertainty to reach correct answers. Using the text-only AgentClinic-MedQA and AgentClinic-MIMIC-IV environments (with GPT-4o for the patient agent and measurement), MedGemma 27B showed superior performance to Gemma 3 27B, exceeding performance of human physicians on AgentClinic-MedQA, and approaching performance of much larger models on both datasets. Results are summarized in Table 11. Of note, the 4B variants (MedGemma and Gemma 3) were not well suited for this task, demonstrating difficulty following system instructions for the agentic framework as provided by AgentClinic.

General purpose benchmarks: Given the limitations and poor performance that many medically-specialized models exhibit when faced with non-medical tasks, we also evaluated for possible tradeoffs

Prompt

You are an expert radiologist. Please succinctly describe the findings for the above chest x-ray.

**MedGemma**

The chest X-ray shows a large left pneumothorax with a chest tube in place. There is also a possible nodule in the right upper lobe. The mediastinum is midline. The heart size is normal. There is no evidence of pulmonary edema.

Reference Information: Radiologist Impression

Moderately large left hydropneumothorax (predominantly air-containing) with incompletely expanded left lung. If persistent, consider ex vacuo pneumothorax/trapped lung condition. Multiple right lung nodules, highly suspicious for metastases.

Prompt

Describe this image and provide the most likely condition. Keep your answer brief

**MedGemma**

The chest X-ray shows a significant opacity in the left lower lung field, obscuring the costophrenic angle. This suggests a possible pleural effusion or consolidation. The right lung appears relatively clear. The heart size is within normal limits. The mediastinum is unremarkable.

Reference Information: Radiologist Impression

Lower left lung consolidation. Small-medium left pleural effusion. Similar though lesser findings right side.

Prompt

What do you think is the most likely condition in this image? Keep your answer brief

**MedGemma**

Based on the image, the most likely condition is a basal cell carcinoma (BCC). The presence of a pearly or waxy bump with visible blood vessels is a classic sign.

Prompt

What next steps would you suggest? keep your answer brief

MedGemma

Consult a dermatologist for a definitive diagnosis and treatment plan.

Reference Information:

Dermatologist suspicion BCC

Prompt

Describe this image and provide the most likely condition. Keep your answer brief

**MedGemma**

The image shows a close-up of a leg with multiple small, red, raised bumps. These are likely insect bites, possibly from mosquitoes, fleas, or other biting insects.

Prompt

What other information would you want to know from the patient to improve your diagnosis? Keep your answer brief

MedGemma

I would ask:

- When did the bites start?
- Are there any other bites on the body?
- Is there any itching or pain associated with the bites?
- Have you been in any areas where insects are common?
- Do you have any known allergies?

Reference Information:

Dermatologist suspicion insect bite

Figure 2 | Example of CXR and dermatology image dialogue via open-ended question answering with MedGemma 4B.

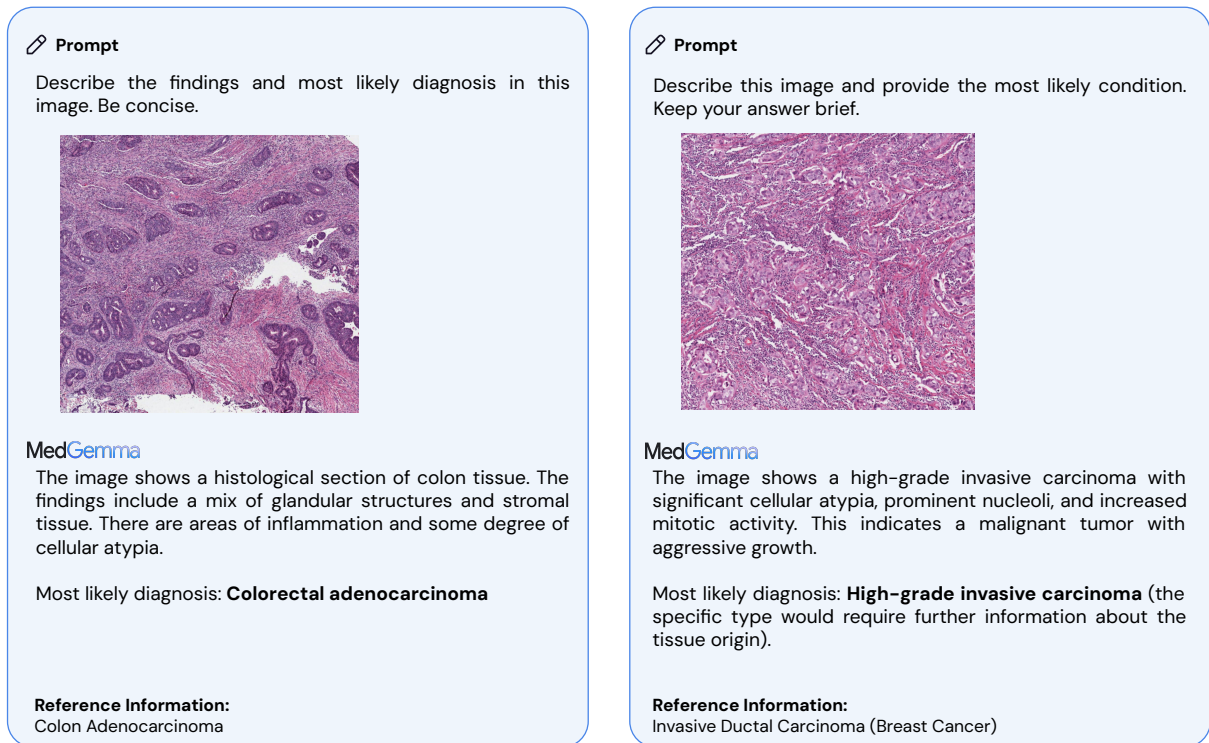


Figure 3 | Example of histopathology open-ended question answering with MedGemma 4B.

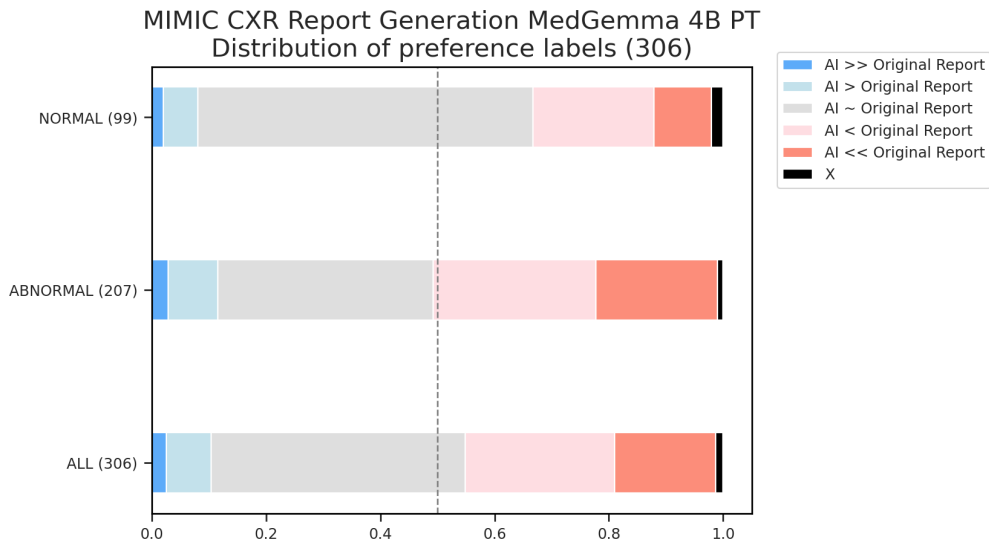


Figure 4 | The MedGemma 4B PT model was used to generate radiology reports on the MIMIC-CXR test set. A single board-certified thoracic radiologist reviewed the MIMIC-CXR report, generated report, and the corresponding CXR image to judge the quality of the reports. Images were reviewed using the original DICOMs on a clinical diagnostic viewer. Across all cases, 81% of MedGemma's CXR reports resulted in the same or superior clinical decision in comparison to the original reports.

Table 7 | Medical image classification with zero-shot generative output for chest X-ray. This table presents the performance on chest X-ray classification for presence of specific conditions, evaluated as a 0-shot, generative task. For MIMIC-CXR and CheXpert datasets, macro F1 is calculated for atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. For CXR14, macro F1 is calculated for lung opacity, pneumothorax, and fracture.

Dataset	Metric	Small Models			Large Models			SOTA VLM §
		MedGemma 4B	Gemma 3 4B	Gemma 3 27B	Gemini v2.5 Flash	Gemini v2.5 Pro	o3	
MIMIC-CXR Med-Gemini test set*	macro F1 (5 conditions)	88.9	81.2	71.7	81.0	85.8	N/A	90.7 Med-Gemini (Yang et al., 2024)
MIMIC-CXR MAIRA test set		40.5	26.7	25.0	32.2	31.9	N/A	51.6 Med-PaLM M (Tu et al., 2024)
CheXpert (OOD)	macro F1 (5 conditions)	48.1	32.6	26.2	37.4	37.0	40.9	49.0 † RadVLM (Deperrois et al., 2025)
CXR14 (OOD)*	macro F1 (3 conditions)	50.1	32.0	31.4	36.6	39.2	32.0	46.7 Med-Gemini (Yang et al., 2024)

* Radiologist adjudicated labels are used (Yang et al., 2024) Section A.1.1., so metrics may not be directly comparable to those reported in the literature for external models.

† RadVLM macro F1 was calculated based on the individual F1 scores reported, estimated from the associated bar chart

§ Prior reported results

Table 8 | Medical image classification with zero-shot generative output for histopathology, dermatology, and retina. Performance on internal classification tasks for histopathology, dermatology, and retinal fundus images; formulated as zero-shot, multiple choice tasks for a generative model.

Modality	Dataset	Metric	Small Models			Large Models	
			MedGemma 4B	Gemma 3 4B	Gemma 3 27B	Gemini v2.5 Flash	Gemini v2.5 Pro
Histopath.	PathMCQA	accuracy	69.8	37.1	42.2	46.9	42.7
Skin	US-Derm MCQA	accuracy	71.8	52.5	66.9	78.4	81.0
Retina	EyePACS	accuracy	64.9	14.4	20.3	17.5	27.7

of specialization. Models were evaluated on MMLU Pro, Global MMLU Lite, and MMMU benchmark datasets, with performance comparisons between MedGemma and Gemma 3 variants shown in (Table 12). MedGemma demonstrated only minor decreases in performance relative to the general models of the same size, further suggesting potential utility for downstream applications that require both specialized as well as generalized capabilities such as instruction following or diverse user interactions.

Observed themes: Several key themes emerged from our analysis. While larger, more computationally expensive models generally performed better, medically-specialized models like MedGemma exhibited a distinct advantage relative to their parameter size. This size-to-performance benefit was especially pronounced in medical vision-based tasks, where smaller clinical models, in some instances, even surpassed the performance of significantly larger general-purpose models. Notably, there was a 500-fold difference in computational cost between MedGemma 4B and the most expensive comparator model, a crucial consideration for practical application development where both development and compute usage are constraints.

We also observed that performance on older, established benchmarks tended to improve with newer models. While this observation reflects genuine advancements in model capabilities, it also raises the possibility of test data leakage, as these benchmarks are publicly available and frequently used in model development and evaluation.

Table 9 | Medical VQA results for SLAKE and VQA-RAD. We include comparisons to other generative models. Additionally, the original VQA-RAD test set includes some duplicated images in the train and test sets (with different questions). As such we have previously described our own splits to avoid this contamination, but precludes direct comparison to externally reported metrics for this dataset.

MedXpertQA Model	SLAKE (English)			VQA-RAD Test split from (Xu et al., 2023)	
	overall token f1	open-ended token recall	closed-ended accuracy	overall token f1	closed Q&A accuracy
Small Models					
MedGemma 4B	72.3	63.3	87.6	49.9	69.1
Gemma 3 4B	40.2	33.3	53.0	33.6	48.7
Gemma 3 27B	42.5	30.8	64.5	42.7	59.4
BiomedGPT-B (Zhang et al., 2023) [§]	85.2	-	89.9	*	*
LLaVA-Med (BioMedCLIP) (Li et al., 2023) [§]	-	87.1	86.8	*	*
Large Models					
Med-Gemini (Yang et al., 2024)	75.8	72.2	84.6	50.1	69.7
Gemini 2.5 Flash	54.6	42.2	80.9	53.6	70.2
Gemini 2.5 Pro	53.1	40.4	78.6	54.2	71.4
o3	55.5	45.0	76.3	52.5	71.9

* Reported results not available for the same test split (only for the original VQA-RAD splits).

- Results for this metric not available in associated report.

§ Prior reported results

Table 10 | 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 section. MedGemma addresses 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. Note that reported metrics across all models may not be directly comparable due to differences in exclusion and inclusion criteria for the test set in the respective citations.

Model	Sections	Clinical Metric (%)
		RadGraph F1 [§]
CXR-RePaiR (Endo et al., 2021)	F only	9.1
\mathcal{M}^2 Transformer (Miura et al., 2020)	F only	22.0
Med-PaLM M, 12B (Tu et al., 2024)	F only	25.2
Med-PaLM M, 84B (Tu et al., 2024)	F only	26.7
MAIRA-1 (Hyland et al., 2023)	F only	24.3
MAIRA-2 (Bannur et al., 2024)	F only	34.6
R2Gen (Chen et al., 2020b)	F + I	13.4
WCT (Yan et al., 2021)	F + I	14.3
CvT-21DistillGPT2 (Nicolson et al., 2023)	F + I	15.4
Flamingo-CXR (Tanno et al., 2024)	F + I	20.5
Med-Gemini-2D (Yang et al., 2024)	F + I	24.4
PaliGemma 2 10B (Steiner et al., 2024)	F + I	29.5
MedVersa (Zhou et al., 2024)	F + I	30.0
MedGemma 4B PT	F + I	29.5

§ Prior reported results except for MedGemma

5. MedGemma Fine-tuning Demonstration

While the MedGemma models achieved strong baseline performance, users who seek to improve performance in specialized scenarios (e.g. conforming to a specific reporting style, classification with

Table 11 | Performance on AgentClinic benchmarks.

Model	AgentClinic-MedQA Accuracy(%)	AgentClinic-MIMIC (OOD) Accuracy(%)
Small Models		
MedGemma 27B	56.2	46.0
Gemma 3 27B	50.7	35.2
Large Models		
DeepSeek R1 (DeepSeek-AI, 2025)	58.1	43.8
Gemini 2.5 Pro	58.3	48.4
o3	65.8	50.6
Human physician	54.0 [*]	N/A

^{*} Human metric based on GPT-4 patient agent as reported in [Schmidgall et al. 2024](#). All other metrics were recalculated using GPT-4o as patient agent to minimize deviation from original AgentClinic report.

Table 12 | Accuracy results on general, non-medical benchmarks.

Type	Benchmark	MedGemma 4B	Gemma 3 4B [§]	MedGemma 27B	Gemma 3 27B [§]
Text-only	MMLU Pro	39.1	43.6	60.2	67.5
	Global MMLU Lite	55.5	54.5	74.5	75.1
Multi-modal	MMMU (val)	47.3	48.8	N/A	64.9

[§] Prior reported results

classes difficult to describe using language alone, new domains that the base models haven’t been trained on) will need to further adapt the models.

We conducted four fine-tuning experiments to demonstrate MedGemma’s capacity to be adapted. There are three multimodal tasks, (1) MIMIC-CXR clinical report generation, (2) SIIM-ARC CXR pneumothorax classification ([Zawacki et al., 2019](#)), and (3) CRC100k histopathology patch classification ([Kather et al., 2018](#)), as well as one text-only task involving long-context electronic health record (EHR) question answering. In these experiments, we adapted the released models via supervised fine-tuning (SFT) for the multimodal tasks and RL for the EHR task. We note that the only task that strictly required the usage of SFT instead of RL is MIMIC-CXR report generation, in order to capture the implicit reporting style; other tasks can be fine-tuned through either SFT or RL. In our experience, the performance trade-off between these two methods needs to be established on a case-by-case basis given the dataset size, distribution shift compared to the original training datasets, and difficulty of the task.

5.1. Fine-tuning for multimodal tasks

We utilized the instruction-tuned version of MedGemma 4B as the base model for the multimodal fine-tuning tasks. For MIMIC-CXR report generation tasks, the image and the indication section of the report were used as the input, and the model was trained to complete the finding and impression sections of the report. For pneumothorax classification and histopathology patch classification, the classification task was formulated as a multiple choice question, the image and a multiple choice question prompt were used as the input, and the model was trained to produce the correct choice. The model was fine-tuned through SFT which involves full parameter updates by optimizing a standard cross-entropy loss function for next token prediction. The input image and prompt were excluded from the loss computation. For SIIM-ARC and CRC100k, since these data sets provide no validation split, we reserved a random 10% per label of the original training data to form a validation set. For each task, we conducted a hyperparameter sweep across three learning rates (1e−7, 5e−7, 1e−6) and

Table 13 | MedGemma fine-tuning results.

Task	Dataset	Metric	MedGemma 4B		SOTA [§]
			Out-of-box	Fine-tuned	
CXR report generation	MIMIC-CXR	RadGraph F1	29.5	30.3	30.0 MedVersa (Zhou et al., 2024)
CXR binary classification	SIIM-ACR Pneumothorax (OOD)	Accuracy	85.9	87.8	88.9 Unichest FT (Dai et al., 2024)
		F1	59.7	71.5	72.5 Unichest FT (Dai et al., 2024)
Histopathology classification	CRC100k (OOD)	Weighted F1	32.8	94.5	97.3 Virchow (Vorontsov et al., 2023)

§ Prior reported results

fine-tuned the model for a single epoch. The checkpoints that achieved the highest performance on the validation set were selected for the final evaluation on the test set.

The results in Table 13 show the effectiveness of further adapting the MedGemma models to specific use cases. The adapted models more closely approached SOTA models on these tasks (Chen et al., 2020a; Dai et al., 2024; Zhou et al., 2024). Notably, the fine-tuned MedGemma 4B established a new SOTA performance with a RadGraph F1 score of 30.3 on the MIMIC-CXR report generation task.

5.2. Fine-tuning for EHR information retrieval and reasoning

While numerous benchmarks assess the ability of language models to reason over EHRs, including emrQA (Pampari et al., 2018), emrKBQA (Raghavan et al., 2021), EHRSQL (Lee et al., 2024), EHRNoteQA (Kweon et al., 2024), and MedAlign (Fleming et al., 2024), we developed a new benchmark for MedGemma specifically focused on longitudinal, outpatient EHR data. Our benchmark, which we call EHRQA, is based on a programmatic framework to generate question-answer (QA) pairs from synthetic, FHIR-formatted records. These records were produced using Synthea (Walonoski et al., 2018), a tool that simulates complete patient medical histories based on population statistics and disease progression models. From an initial 100 synthetic records, we selected 81 that fit within a 32k context window, splitting them into training (42), validation (20), and test (19) sets. Each patient record is comprehensive, containing hundreds to thousands of FHIR entries across various resource types like Conditions, Medications, and Observations. A key limitation of this dataset, however, is the absence of clinical notes, which we hope to address in future work.

Our QA generation framework treated each patient’s FHIR record as a collection of ground truth facts. It expanded upon these facts through a multi-hop reasoning process (Yang et al., 2018) that integrated information from external medical knowledge bases (SNOMED, RxNorm, LOINC), applied programmatic reasoning (e.g., temporal, arithmetic), and leveraged inter-dependencies within the patient’s data programmatically. These derived facts are then converted into natural language QA pairs using templated questions and large language model (LLM) rephrasing. In collaboration with clinicians, we developed 42 distinct question types, grouped into 10 categories, to reflect queries from both healthcare professionals and consumers (see Appendix Table A8 for more details). This process generated approximately 200 questions per patient. In total, there are 10,437 QA examples in the training split, 5,133 QA examples in the validation split, and 4,377 QA examples in the test split. All questions were designed for automated evaluation (e.g., regex matching) and support multiple-choice, Yes/No, and simple free-response formats. For evaluation, the model is prompted with a condensed, plain-text representation of the patient’s FHIR data and the corresponding question. Final accuracies are calculated by averaging the scores across the 10 question categories. The prompts used for rephrasing and evaluation are available in Appendix Table A9.

Our initial results revealed a performance gap between MedGemma 27B and larger models. To address this, we fine-tuned MedGemma 27B using RL on the EHRQA training set. Applying RL to the MedGemma 27B resulted in a substantial improvement in its EHRQA accuracy, closing the gap with top-performing models, as shown in Table 14. Notably, the greatest gains were observed in question categories requiring reasoning across inter-dependent records (Appendix Figure A3). This experiment underscores the potential of fine-tuning smaller, specialized models like MedGemma for effective use in EHR applications.

Table 14 | EHRQA accuracy results (OOD)

Type	Small Models			Large Models		
	MedGemma 27B Text-only	Gemma 3 27B	MedGemma 27B Text-only (RL)	Gemini 2.5 Flash	Gemini 2.5 Pro	o3
EHRQA	86.3	84.2	93.6	95.0	95.4	92.5

6. MedSigLIP Evaluations

MedSigLIP was evaluated both on zero-shot classification performance and linear probe classification performance (also referred to as data-efficient classification). Zero-shot classification performance was assessed as a measure of baseline performance, and linear probe performance was assessed to gauge performance on target conditions after additional training with logistic regression.

Evaluation datasets spanned four modalities: chest X-ray, dermatology, ophthalmology and histopathology. For chest X-ray, we followed the evaluation datasets and framework in ELIXR (Xu et al., 2023). For linear probing, we utilized the same data from CheXpert and CXR14 and evaluated on seven findings (atelectasis, cardiomegaly, airspace opacity, fracture, pneumothorax, consolidation, pleural effusion, and pulmonary edema). For zero-shot, we compared with ELIXR on the 13 positive findings from CheXpert test set. For dermatology, we evaluated on US-Derm MCQA. The US-Derm MCQA test set has 79 dermatological conditions that overlap with the classification labels from the training dataset, thus we pick this 79-condition subset (1612 patients) for our vision based classification and zero-shot evaluation. For ophthalmology, we used the EyePACS test dataset. For histopathology, patches were extracted from whole slide images across a variety of tissue types and tasks as in Yang et al. (2024) and these patches were treated as individual images. See Appendix Table A2 for a description of the detailed classes for the evaluations.

Zero-shot evaluation approach and metrics: For each class within each condition, one or more text prompts were used to represent the class. When there were multiple prompts for the same class, the text embeddings for each prompt were averaged together to obtain a single embedding per class. Cosine similarity was then calculated between the image embedding and text embeddings for the class options, softmax was applied to output scores, and the AUC (area under receiver operating characteristic curve) was calculated. In the case of multiclass conditions, 1-vs-all AUC was calculated. For dermatology, zero-shot prompts are directly their condition names. Prompts for ophthalmology, chest X-ray, histopathology are provided in Appendix Tables A3, A4, and A5.

Linear probe evaluation approach and metrics: Linear probe (data efficient) classification was assessed by extracting image embeddings from MedSigLIP (without use of the text encoder) and training a logistic regression using the SAGA solver (Defazio et al., 2014) on the train set embeddings, with the hyper-parameters chosen on the validation set, then evaluating on the test set.

7. MedSigLIP Results

Zero-shot and linear probe classification results and evaluation tasks for MedSigLIP and corresponding Health AI Developer Foundations (HAI-DEF) (Kiraly et al., 2024) models are summarized below. For the HAI-DEF Derm Foundation and Path Foundation models, zero-shot classification was not possible with the comparator models because of their image-only nature.

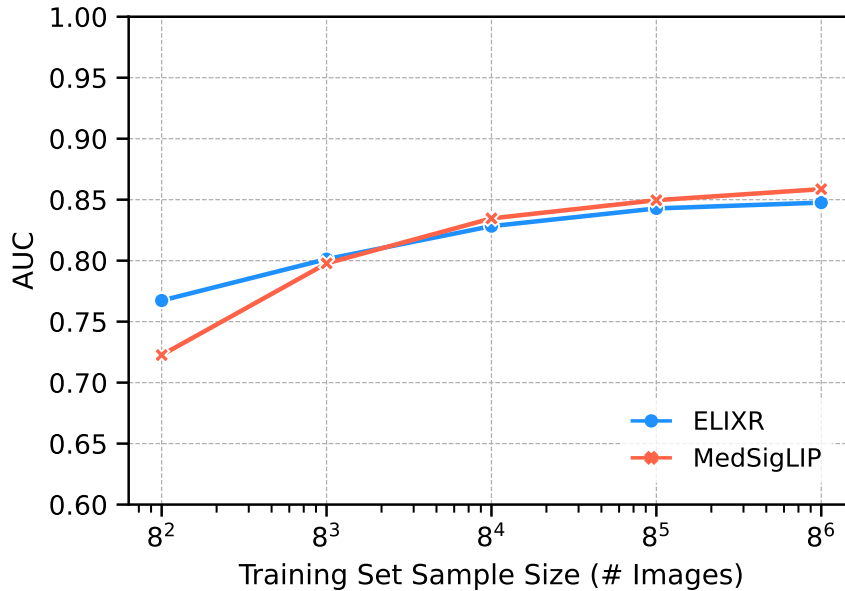


Figure 5 | Average results for data efficient learning on 7 Chest X-ray findings on CheXpert and CXR14 datasets compared to HAI-DEF's CXR Foundation model based on ELIXR (Xu et al., 2023). Individual results per condition per dataset can be found in Appendix Figures A1 and A2.

On CXR, MedSigLIP is compared to the HAI-DEF CXR Foundation model, which is based on ELIXR (Xu et al., 2023), with zero-shot results shown in Table 16 and linear probe classification results shown in Figure 5. On average, MedSigLIP's zero-shot CXR performance was 2.0% higher than CXR Foundation despite MedSigLIP's lower image resolution (448×448 vs 1280×1280) and multi-domain expertise, suggesting that MedSigLIP can serve as a strong foundation model. Notably, classification of fractures, which has historically proven difficult, improved over ELIXR by 7.1%. Across the 7 findings used for linear probing, MedSigLIP also demonstrated strong performance when training set sample size is larger or equal to 512 examples, Figure 5.

Dermatology, ophthalmology and histopathology results are summarized in Table 15. For dermatology, both MedSigLIP zero-shot and linear probes outperformed linear probes with Derm Foundation on the task of distinguishing between 79 skin conditions. For ophthalmology, on the 5-class task of classifying diabetic retinopathy (none, mild, moderate, severe, and proliferative), linear probe performance exceeded zero-shot performance by 9.8%, though no HAI-DEF model currently exists for comparison. For histopathology, linear probing achieved a modest improvement over zero-shot classification using MedSigLIP, and performing close to Path Foundation which has linear probe performance that is 1.9% better on average.

MedSigLIP offers strong baseline performance across a variety of medical domains as a single model and generally does well with zero-shot classification even when compared to linear probes with dedicated, domain-specific models. Further training with task-specific images and logistic regression can yield stronger results.

Table 15 | AUCs for dermatology, ophthalmology, and histopathology findings with MedSigLIP and HAI-DEF image models (Kiraly et al., 2024).

Domain	Finding	N	No. Classes	Resolution	MedSigLIP		HAI-DEF (image) [§]	
					Zero-Shot	Linear Probe	Resolution	Linear Probe
Dermatology	Skin Conditions	1612	79	448x448	0.851	0.881	448x448	0.843
Ophthalmology	Diabetic Retinopathy	3161	5	448x448	0.759	0.857	N/A	N/A
Histopathology	Invasive Breast Cancer	5000	3	448x448	0.933	0.930	224x224	0.943
	Breast NP	5000	3		0.721	0.727		0.758
	Breast TF	5000	3		0.780	0.790		0.832
	Cervical Dysplasia	5000	3		0.889	0.864		0.898
	Prostate Cancer Needle Core Biopsy	5000	4		0.892	0.886		0.915
	Radical Prostatectomy	5000	4		0.896	0.887		0.921
	TCGA Study Types	5000	10		0.922	0.970		0.964
	Tissue Types	5000	16		0.930	0.972		0.947
Average					0.870	0.878		0.897

§ Prior reported results. HAI-DEF (image) refers to the image-only foundation models in HAI-DEF: Derm Foundation and Path Foundation models while MedSigLIP is a single model.

Table 16 | Zero-shot AUCs for chest X-ray findings with MedSigLIP and HAI-DEF's chest X-ray foundation model based on ELIXR (Xu et al., 2023).

Domain	Finding	N	No. Classes	MedSigLIP		HAI-DEF (image) [§]	
				Resolution	Zero-Shot	Resolution	Zero-Shot
CXR	Enlarged Cardiomeastinum	518	2	448x448	0.858	1280x1280	0.800
	Cardiomegaly				0.904		0.891
	Lung Opacity				0.931		0.888
	Lung Lesion				0.822		0.747
	Consolidation				0.880		0.875
	Edema				0.891		0.880
	Pneumonia				0.864		0.881
	Atelectasis				0.836		0.754
	Pneumothorax				0.862		0.800
	Pleural Effusion				0.914		0.930
	Pleural Other				0.650		0.729
	Fracture				0.708		0.637
	Support Devices				0.852		0.894
Average					0.844		0.824

§ Prior reported results

8. Discussion

We introduced MedGemma, a new collection of medical vision-language foundation models and MedSigLIP, a multi-domain medical image encoder. These models were built upon Gemma 3, with optimization for medical domains. We evaluated across a range of medical benchmarks across clinical reasoning, biomedical knowledge, report generation, and medical image classification, finding strong performance for MedGemma and MedSigLIP. Performance improved further after fine-tuning, highlighting the potential for these open models to be used as a starting point for developing useful AI applications for healthcare.

With an increasing number of options available to developers building AI applications in healthcare, MedGemma provides specific advantages over general models. These advantages are largely due to optimized incorporation of domain specific data for both pre-training and post-training and are illustrated by the improvements over base Gemma 3 models across all benchmarks evaluated and the achievement of performance on par with much larger models.

When compared to general API-based models like Gemini, MedGemma is likely the preferred model if the use case requires any of the following: a frozen model for documentation and reliability, sensitivity to training or inference costs, ability to run locally or offline, specific medical image and text capabilities, or full control over model adaptation. Large models like Gemini remain a viable choice where the user requires optimal broad performance without the above constraints, and large models may additionally be used in concert with models like MedGemma in agentic settings.

The MedGemma collection of models enables a wide range of potential downstream applications for the developer community. The multimodal capabilities, including access to image and text embeddings, may be particularly useful for medical image retrieval. This could aid in interpretation by referencing similar past cases as well as enabling development of research cohorts and creating educational tools. MedGemma allows for the integration of diverse data, linking radiology, histopathology, ophthalmology, and dermatology images with clinical information. The specialized text capabilities of the models can also extract key concepts from imaging reports and clinical notes, streamlining tasks such as matching patients for clinical trials, conducting pharmacovigilance reviews, or analyzing healthcare quality metrics. The models' ability to understand medical images and generate reports can also be fine-tuned to better assist radiologists and other clinicians in their workflow and improve how findings are communicated to patients. In addition to standalone use, these models can also serve as powerful tools within agentic frameworks, combining abilities across different modalities for customized and comprehensive solutions.

In this report, we evaluated the performance of MedGemma and MedSigLIP on a broad set of established benchmarks in order to provide a snapshot of the model capabilities. However, we note that limitations exist for these benchmarks. For one, automated benchmarks represent only the first step towards validating real-world utility ([Alaa et al., 2025](#); [Mahmood, 2025](#)). Additionally, some benchmarks may be near saturation in terms of model performance, with minimal headroom for improvements, thus hindering the measurement of progress. As such, further work is warranted to continue evaluation of these models on new, high quality (and more challenging) benchmarks aimed at better reflecting real-world utility (e.g. [Bedi et al., 2025](#)). More work is also needed to understand the performance capabilities and requirements in regard to actual application development, including their incorporation into agentic frameworks. These efforts will inform optimal use cases as well as development of future model versions with extended capabilities.

We openly released MedGemma and MedSigLIP to facilitate their widespread evaluation, improvement, and adaptation by the community. Openness is critical in many healthcare applications, as it provides developers with predictability and the flexibility for extensive model adaptation and evaluation. We hope that our approach accelerates the development of AI applications across a broad array of healthcare use cases.

9. Conclusion

In this work, we showed that MedGemma models demonstrate robust capabilities across a variety of vision-language and text-only medical tasks. We also showed that MedSigLIP demonstrates robust multi-domain capabilities, and can thus serve as a strong medical foundation model. The breadth and efficiency of these models offers exciting possibilities to address a range of use cases. At the same time, thoughtful validation of safety, performance, and reliability for any downstream applications remains a critical aspect to advance the use of multimodal AI models in medicine. By providing these MedGemma and MedSigLIP models to the developer community with a permissive license, we hope to see them enable useful and innovative medical applications.

10. Model availability

The models have been released openly at the main Google Health AI Developer Foundations site at <https://goo.gle/hai-def>. Further details specifically about the MedGemma collection of models can be found at <https://goo.gle/medgemma>.

11. Contributions and Acknowledgments

Contributions

Technical Leads

Lin Yang[†]
Andrew Sellergren^{*}
Sahar Kazemzadeh^{*}
Fereshteh Mahvar

Core contributors

Tiam Jaroensri
Atilla Kiraly
Madeleine Traverse
Timo Kohlberger
Shawn Xu
Fayaz Jamil
Cían Hughes
Charles Lau
Justin Chen
Liron Yatziv
Tiffany Chen
Bram Sterling
Stefanie Anna Baby
Susanna Maria Baby
Jeremy Lai
Samuel Schmidgall
Lu Yang
Kejia Chen
Per Bjornsson
Shashir Reddy
Ryan Brush
Kenneth Philbrick
Mercy Asiedu
Ines Mezerreg
Howard Hu
Howard Yang
Richa Tiwari
Sunny Jansen
Preeti Singh
Yun Liu
Shekoofeh Azizi

Contributors

Aishwarya Kamath
Johan Ferret
Shreya Pathak
Nino Vieillard
Ramona Merhej
Sarah Perrin
Tatiana Matejovicova
Alexandre Ramé
Morgane Riviere
Louis Rouillard
Thomas Mesnard
Geoffrey Cideron
Jean-bastien Grill

Sabela Ramos
Edouard Yvinec
Michelle Casbon
Elena Buchatskaya
Jean-Baptiste Alayrac
Dmitry (Dima) Lepikhin
Vlad Feinberg
Sebastian Borgeaud
Alek Andreev
Cassidy Hardin
Robert Dadashi
Léonard Hussenot
Armand Joulin
Olivier Bachem

Sponsors

Yossi Matias
Katherine Chou
Avinatan Hassidim
Kavi Goel
Clement Farabet
Joelle Barral
Tris Warkentin
Jonathon Shlens
David Fleet

Launch Support

Victor Cotruta
Omar Sanseviero
Gus Martins
Phoebe Kirk
Anand Rao

Leads

Shravya Shetty
David F. Steiner
Can Kirmizibayrak
Rory Pilgrim[†]
Daniel Golden[†]

^{*} Co-first author

[†] Co-last author

Acknowledgements

Many teams from both Google Research and Google DeepMind collaborated extensively on this project. We thank Ellery Wulczyn and Dale Webster for their feedback and insight, which significantly enhanced this report. We thank Naama Hammel, Liam Foster, and Kapil Parakh for their review of the qualitative examples shown in this manuscript. We thank Andreas Steiner and Xiao Wang for sharing their expertise with the Big Vision infrastructure. The results shown here are in part based upon data generated by the [TCGA Research Network](#).

Use of AI in Manuscript Preparation

The introduction section of this manuscript was drafted manually and then further refined using Gemini 2.5 Pro. Final manual checks were performed to ensure content accuracy. The authors take full responsibility for the content.

References

- Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12, 2017.
- 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.
- Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.
- 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.
- Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938, 2024.
- Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>, 2024.
- Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020a. URL <https://arxiv.org/abs/2006.10029>.
- 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*, 2020b.
- 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.
- Tianjie Dai, Ruipeng Zhang, Feng Hong, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Unichest: Conquer-and-divide pre-training for multi-source chest x-ray classification, 2024. URL <https://arxiv.org/abs/2312.11038>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

- Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhertz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, et al. Radvlm: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333*, 2025.
- 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.
- Endurance Evbayekha, Abiodun Benjamin Idowu, and Shane LaRue. Sacubitril/valsartan vs ace inhibitors or arbs: A systematic review and meta-analysis of randomized trials. *JACC: Advances*, 4(3):101598, 2025.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030, Mar. 2024. doi: 10.1609/aaai.v38i20.30205. URL <https://ojs.aaai.org/index.php/AAAI/article/view/30205>.
- Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- 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.
- 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.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- 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.
- IQVIA. Introducing iqvia medical reasoning (med-r1 8b): Best-in-class medical reasoning llm. Web Page, April 2025. URL <https://www.iqvia.com/blogs/2025/04/introducing-iqvia-medical-reasoning-med-r1-8b>. Accessed 2025-06-11.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, 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.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- John Snow Labs. Jsl-medllama-3-8b-v2.0. <https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0>, 2024. Accessed 2025-06-11.
- 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.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue, 2018. URL <https://doi.org/10.5281/zenodo.1214456>. [Data set].
- Atilla P Kiraly, Sebastien Baur, Kenneth Philbrick, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Nick George, Fayaz Jamil, Jing Tang, et al. Health ai developer foundations. *arXiv preprint arXiv:2411.15128*, 2024.
- Sunjun Kweon, Jiyouon Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. PhysioNet, version 1.0.1, June 2024. URL <https://doi.org/10.13026/acga-ht95>.
- 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.
- 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.
- Gyubok Lee, Sunjun Kweon, Seongsu Bae, and Edward Choi. Overview of the EHRSQL 2024 shared task on reliable text-to-SQL modeling on electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 644–654. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.clinicalnlp-1.62. URL <https://aclanthology.org/2024.clinicalnlp-1.62/>.

- 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*, 36:28541–28564, 2023.
- 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, 2024.
- Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.
- Faisal Mahmood. A benchmarking crisis in biomedical machine learning. *Nature Medicine*, 31(4):1060–1060, 2025.
- 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.
- 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.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023.
- Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, et al. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*, 2024.
- OpenAI. GPT-4V(ision) Technical Work and Authors. Technical report, OpenAI, 2023. URL <https://cdn.openai.com/contributions/gpt-4v.pdf>.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368. Association for Computational Linguistics, october-november 2018. doi: 10.18653/v1/D18-1258. URL <https://aclanthology.org/D18-1258/>.
- Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. emrK-BQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73. Association for Computational Linguistics, june 2021. doi: 10.18653/v1/2021.bionlp-1.7. URL <https://aclanthology.org/2021.bionlp-1.7/>.
- 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.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.
- 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, 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, 2023b.
- 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*, 2023c.
- Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.
- 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.
- 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.
- 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.

- Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.
- Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018. doi: 10.1093/jamia/ocx079.
- 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.
- Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *Physio Net*, 2021.
- Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.
- 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.
- Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Attila Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, oct – nov 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259/>.
- Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation>, 2019. Kaggle.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.
- Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pages arXiv–2305, 2023.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation, 2024. URL <https://arxiv.org/abs/2405.07988>.

Appendix

A. Manual Evaluation of Radiology reports

Table A1 provides the detailed rubric definitions provided to the radiologist for scoring the generated CXR reports.

Table A1 | Human evaluation rubric comparing AI generated radiology reports to original reports.

Rubric Score	Rubric Definition
AI » Original Report	AI report captures key clinically relevant findings that are not found in original report. AI report would result in correct patient management and original report would not.
AI > Original Report	AI report captures more relevant findings, but both would result in the same correct patient management.
AI ~ Original Report	Both reports capture similar findings in the image and would result in correct patient management.
AI < Original Report	Original report captures more relevant findings, but both would result in the same correct patient management.
AI « Original Report	Original report captures key clinically relevant findings that are not found in AI report. original report would result in correct patient management and AI report would not.
X	Neither report would result in correct patient management.

B. Evaluation Prompts

For evaluating MedSigLIP zero-shot tasks, Tables A2 to A5 list the prompts used.

For evaluating MedGemma, the default temperature of 0.0 was used. All other models made use of their default system temperature. Two different system prompts were used in providing results for other LLMs directly evaluated:

- *You are a helpful radiology assistant.* For radiology-based tasks.
- *You are a helpful medical assistant.* For all other medical tasks.

Per-dataset prompts used are shown in Appendix Tables A6, A7.

Table A2 | MedSigLIP Class Definitions

Classes for the findings that MedSigLIP was evaluated on.

Domain	Finding	Num Classes	Class Names
Radiology	CXR	2	Positive/negative for Cardiomegaly, Consolidation, Edema, Enlarged cardiomeastinum, Fracture, Lung lesion, Lung opacity, Pleural effusion, Pleural other, Pneumonia, Pneumothorax, Support devices.
Histopathology	Invasive Breast Cancer	3	Benign, Invasive Carcinoma, DCIS
	Breast NP	3	NP1, NP2, NP3
	Breast TF	3	TF1, TF2, TF3
	Cervical Dysplasia	3	Normal, CIN Grade 1, CIN Grade 2+
	Prostate Cancer Needle Core Biopsy	4	Benign, GP3, GP4, GP5
	Radical Prostatectomy	4	Benign, GP3, GP4, GP5
	TCGA Study Types	10	BLCA, BRCA, COAD, HNSC, KIRC, LIHC, LUAD, LUSC, OV, STAD
	Tissue Types	16	Appendix, Breast, Cervix, Colon, Fallopian Tube, Gallbladder, Liver, Lymph node, Ovary, Placenta, Prostate, Skin, Thyroid, Upper GI, Uterus, Vas deferens
Dermatology	Skin Conditions	79	Acanthosis nigricans, Acne, Acne keloidalis, Actinic Keratosis, Allergic Contact Dermatitis, Alopecia Areata, Amyloidosis of skin, Androgenetic Alopecia, Angiokeratoma of skin, Atypical Nevus, Basal Cell Carcinoma, Bullous Pemphigoid, Burn of skin, Candida, Clavus, Comedone, Condyloma acuminatum, Cutaneous T Cell Lymphoma, Cutaneous lupus, Cutaneous sarcoidosis, Cyst, Dermatofibroma, Drug Rash, Eczema, Erythema ab igne, Erythema multiforme, Folliculitis, Folliculitis decalvans, Fordyce spots, Granuloma annulare, Hemangioma, Herpes Zoster, Hidradenitis, Idiopathic guttate hypomelanosis, Infected skin lesions, Inflicted skin lesions, Insect Bite, Intertrigo, Irritant Contact Dermatitis, Keratosis pilaris, Knuckle pads, Lentigo, Lichen Simplex Chronicus, Lichen planopilaris, Lichen planus/lichenoid eruption, Lichen sclerosus, Lipodermatosclerosis, Lipoma, Livedo reticularis, Melanocytic Nevus, Melanoma, Melasma, Milia, Nevus sebaceous, O/E - ecchymoses present, Onychomycosis, Perioral Dermatitis, Photodermatitis, Pigmented purpuric eruption, Pityriasis lichenoides, Pityriasis rosea, Post-Inflammatory hyperpigmentation, Prurigo nodularis, Psoriasis, Pyoderma Gangrenosum, Pyogenic granuloma, Rosacea, SCC/SCCIS, SK/ISK, Scabies, Scar Condition, Seborrheic Dermatitis, Skin Tag, Stasis Dermatitis, Tinea, Tinea Versicolor, Urticaria, Verruca vulgaris, Vitiligo
Ophthalmology	Diabetic Retinopathy (DR)	5	No DR, Mild DR, Moderate DR, Severe DR, Proliferative DR

Table A3 | Zero shot prompts for ophthalmology

<i>Ophthalmology</i>	
Finding	Severity
Diabetic retinopathy	diabetic retinopathy severity: none.
	diabetic retinopathy severity: mild.
	diabetic retinopathy severity: moderate.
	diabetic retinopathy severity: severe.
	diabetic retinopathy severity: proliferative.

Table A4 | Zero shot prompts for CXR

<i>Chest X-ray</i>		
Finding	Condition Absent	Condition Present
Atelectasis	no atelectasis no acute cardiopulmonary process	adjacent atelectasis bibasilar atelectasis there is atelectasis
Cardiomegaly	heart size is normal cardiac size is within normal limits cardiothoracic ratio within normal limits no acute cardiopulmonary process normal study no evidence of cardiomegaly	mild cardiomegaly moderate cardiomegaly severe cardiomegaly enlarged cardiac silhouette
Consolidation	no focal consolidation no acute cardiopulmonary process normal study	alveolar consolidation densely consolidated lobe consolidation airspace consolidation bibasilar consolidations suggestive of consolidation
Edema	no pulmonary edema no acute cardiopulmonary process normal study	mild pulmonary edema moderate pulmonary edema severe pulmonary edema
Enlarged cardiomediatinum	no acute cardiopulmonary process cardiomediatinal silhouette is normal	widened cardiomediatinum
Fracture	no acute cardiopulmonary process normal study	rib fractures rib fracture
Lung lesion	no acute cardiopulmonary process	lytic lesion cavitary lesion parenchymal lesion
Lung opacity	no focal opacity lung volumes are normal normal study normal lung volumes lungs are clear no focal consolidation no evidence of airspace consolidation no infiltrate no airspace opacity pulmonary parenchyma is clear	bilateral opacities basal opacity opacification is present bibasilar opacities increased opacification consolidative opacity parenchymal opacities airspace opacification cannot be excluded there is airspace opacification
Pleural effusion	no evidence of pleural effusion no acute cardiopulmonary process normal study	left pleural effusion right pleural effusion bilateral pleural effusions
Pleural other	no acute cardiopulmonary process	blunting of costophrenic angle pleural thickening
Pneumonia	lungs are clear no acute cardiopulmonary process	suggestive of pneumonia
Pneumothorax	no pneumothorax no acute cardiopulmonary process normal study	apical pneumothorax
Support devices	no acute cardiopulmonary process	monitoring and support devices NG tube ET tube catheter PIC line

Table A5 | Zero shot prompts for histopathology tasks

Condition	Class	Prompt
Invasive Breast Cancer	1	<ul style="list-style-type: none"> • region of an HE histopathology image showing benign breast tissue • HE-stained image demonstrating normal breast tissue architecture • histopathology image with breast lobules and ducts consistent with benign breast tissue • microscopic view of breast tissue with regular uniformly spaced glands indicative of benign breast tissue • HE histopathology image showing normal breast stroma and absence of cellular atypia characteristic of benign breast tissue
	2	<ul style="list-style-type: none"> • region of an HE histopathology image showing invasive breast carcinoma • HE-stained region demonstrating features of invasive breast carcinoma • histopathology image with areas of dense cellularity and atypical cells consistent with invasive breast carcinoma • microscopic view of breast tissue with disorganized glandular architecture consistent with invasive carcinoma • HE histopathology image with regions of stromal invasion by malignant cells indicating invasive breast carcinoma
	3	<ul style="list-style-type: none"> • region of an HE histopathology image showing ductal carcinoma in situ (DCIS) • HE-stained image demonstrating abnormal cells confined within the breast duct • histopathology image: atypical cells within the duct consistent with DCIS • microscopic view of breast tissue showing abnormal cellular proliferation within a duct consistent with DCIS • HE histopathology section with non-invasive intraductal malignant cells suggestive of DCIS
Breast NP		

(Continued on next page)

(Continued)

Condition	Class	Prompt
	1	<ul style="list-style-type: none"> • HE histopathology image demonstrating invasive breast carcinoma with low-grade nuclear features (nucleopleomorphism score 1) • region of invasive breast carcinoma exhibiting relatively uniform nuclei with inconspicuous nucleoli consistent with a low nucleopleomorphism score • microscopic view of invasive breast carcinoma showing limited variation in nuclear shape and size (nucleopleomorphism score 1) • focus on infiltrating tumor cells within an HE-stained image demonstrating bland nuclear features suggestive of low-grade invasive breast carcinoma • HE histopathology section of invasive breast carcinoma with well-formed glands and minimal nuclear atypia indicating a low nucleopleomorphism score
	2	<ul style="list-style-type: none"> • HE histopathology image demonstrating invasive breast carcinoma with moderate nuclear pleomorphism (score 2) • region of invasive breast carcinoma exhibiting some variation in nuclear size and shape with some visible nucleoli (nucleopleomorphism score 2) • microscopic view of invasive breast carcinoma showing moderate nuclear atypia including enlarged nuclei and prominent nucleoli (nucleopleomorphism score 2) • focus on tumor cells with increased nuclear irregularity compared to normal breast tissue consistent with a nucleopleomorphism score of 2 • HE histopathology section of invasive breast carcinoma with moderately pleomorphic nuclei and discernible nucleoli (nucleopleomorphism score 2)

(Continued on next page)

(Continued)

Condition	Class	Prompt
	3	<ul style="list-style-type: none"> • HE histopathology image demonstrating invasive breast carcinoma with high-grade nuclear features (nucleopleomorphism score 3) • microscopic view of invasive breast carcinoma showing significant nuclear atypia including marked variation in size, shape, and prominent nucleoli (nucleopleomorphism score 3) • region of invasive breast carcinoma exhibiting large nuclei with considerable variation in size and shape as well as large irregular nucleoli (nucleopleomorphism score 3) • HE-stained image highlighting markedly pleomorphic nuclei of tumor cells consistent with high-grade invasive breast carcinoma • focus on an area of invasive breast carcinoma demonstrating significant nuclear abnormalities and variability consistent with a nucleopleomorphism score of 3
Breast TF	1	<ul style="list-style-type: none"> • well-formed tubules are visible within this region of invasive breast carcinoma on an HE histopathology image (tubule formation score 1) • this microscopic view demonstrates a predominance of well-defined glandular structures indicating a tubule formation score of 1 in this breast carcinoma • the infiltrating carcinoma cells display preserved tubular architecture resembling normal breast tissue consistent with a tubule formation score of 1 • a tubule formation score of 1 is evident on this HE-stained image showcasing numerous well-formed tubules within the breast carcinoma • this region of invasive breast carcinoma depicted on an HE histopathology image exhibits a high degree of tubular differentiation (tubule formation score 1)

(Continued on next page)

(Continued)

Condition	Class	Prompt
	2	<ul style="list-style-type: none"> • this HE histopathology image of invasive breast carcinoma shows a moderate degree of tubule formation (score 2) • while some tubular structures are present this microscopic view highlights a less organized pattern of growth within the breast carcinoma aligning with a tubule formation score of 2 • a mix of both well-formed tubules and areas lacking clear glandular structures suggests a tubule formation score of 2 in this breast carcinoma • the presence of some discernible tubules alongside regions of less-defined glandular architecture indicates a tubule formation score of 2 • this invasive breast carcinoma displays an intermediate level of tubular differentiation consistent with a tubule formation score of 2
	3	<ul style="list-style-type: none"> • tubular structures are largely absent within this region of invasive breast carcinoma indicating a tubule formation score of 3 on an HE histopathology image • this breast carcinoma demonstrates disorganized growth with minimal gland formation consistent with a tubule formation score of 3 • poorly differentiated tumor cells lacking discernible tubules predominate in this microscopic view suggestive of a tubule formation score of 3 • a tubule formation score of 3 is evident within this HE-stained image where the invasive carcinoma shows a scarcity of well-defined glandular structures • limited tubule formation characterizes this region of breast carcinoma resulting in a tubule formation score of 3

Cervical Dysplasia*(Continued on next page)*

(Continued)

Condition	Class	Prompt
	1	<ul style="list-style-type: none"> • HE-stained image of a cervical biopsy demonstrating normal squamous epithelium and underlying stroma • microscopic view of a cervical biopsy showing regular stratification of the squamous epithelium without atypia • region of a cervical biopsy with well-defined mature squamous epithelium and unremarkable stroma • cervical biopsy demonstrating normal endocervical glands and adjacent squamous epithelium • HE histopathology of a cervical biopsy showcasing a section of the transformation zone with no pathologic findings
	2	<ul style="list-style-type: none"> • HE-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 • region of a cervical biopsy exhibiting CIN 1 with mild nuclear atypia and koilocytic change (perinuclear halos) • cervical biopsy demonstrating CIN 1 with subtle nuclear abnormalities confined to the lower epithelial layers • HE histopathology of a cervical biopsy with a focus on CIN 1 characterized by slight disorganization of the squamous epithelium

(Continued on next page)

(Continued)

Condition	Class	Prompt
	3	<ul style="list-style-type: none"> • HE-stained image of a cervical biopsy demonstrating high-grade dysplasia with significant nuclear atypia and architectural disarray • microscopic view of a cervical biopsy showcasing high-grade squamous intraepithelial lesion (HSIL) with loss of epithelial maturation and cellular disorganization • region of a cervical biopsy exhibiting high-grade dysplasia characterized by increased mitotic activity, nuclear pleomorphism, and loss of normal epithelial polarity • cervical biopsy demonstrating high-grade dysplasia with marked cellular atypia • HE histopathology of a cervical biopsy focusing on high-grade dysplasia showing prominent nuclear abnormalities and disruption of the normal epithelial architecture

Prostate Cancer Needle Core Biopsy

1

- HE histopathology image showing benign prostate tissue within a prostate specimen
- prostate image demonstrating normal prostatic glands and stroma consistent with benign tissue
- microscopic view of prostate with well-defined regularly spaced glands and intervening stroma indicative of benign prostatic tissue
- focus on a region of benign prostate tissue within a prostate specimen exhibiting a normal glandular architecture and stromal component
- HE section of prostate showcasing benign prostatic glands lined by a bilayer epithelium (basal and luminal cells)

(Continued on next page)

(Continued)

Condition	Class	Prompt
	2	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 3 prostate cancer within a prostate specimen • prostate image highlighting individual well-formed glands characteristic of Gleason pattern 3 carcinoma • microscopic view of a prostate with discrete uniformly sized and shaped glands consistent with Gleason pattern 3 prostate cancer • focus on Gleason pattern 3 prostate cancer within prostate specimen showing relatively distinct glandular structures with minimal variability • HE section of prostate demonstrating areas of Gleason pattern 3 adenocarcinoma marked by separate glands composed of tumor cells infiltrating between benign glands
	3	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 4 prostate cancer within a prostate specimen • prostate image highlighting areas of fused glands or poorly formed glands consistent with Gleason pattern 4 carcinoma • microscopic view of prostate showcasing irregular glandular structures with varying shapes and sizes indicative of Gleason pattern 4 prostate cancer • focus on Gleason pattern 4 prostate cancer within a prostate specimen demonstrating loss of normal glandular architecture and a more disorganized growth pattern • HE section of prostate demonstrating regions of Gleason pattern 4 adenocarcinoma characterized by fused or poorly formed glands

(Continued on next page)

(Continued)

Condition	Class	Prompt
	4	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 5 prostate cancer within a prostate specimen • prostate image highlighting sheets of tumor cells lacking any glandular formation characteristic of Gleason pattern 5 carcinoma • microscopic view of prostate showing solid areas of tumor cells representing Gleason pattern 5 prostate cancer • focus on Gleason pattern 5 prostate cancer within a prostate specimen demonstrating poorly differentiated cells and absence of gland formation • HE section of prostate with regions of Gleason pattern 5 adenocarcinoma characterized by a complete lack of glandular differentiation and presence of solid tumor cell clusters
Radical Prostatectomy	1	<ul style="list-style-type: none"> • HE histopathology image showing benign prostate tissue • small region from prostate specimen demonstrating normal prostatic glands and stroma consistent with benign tissue • microscopic view of a prostate specimen with well-defined regularly spaced glands and intervening stroma indicative of benign prostatic tissue • focus on a region of benign prostate tissue exhibiting a normal glandular architecture and stromal component • HE section showcasing benign prostatic glands lined by a bilayer epithelium (basal and luminal cells)

(Continued on next page)

(Continued)

Condition	Class	Prompt
	2	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 3 prostate cancer • small region of prostate on histopathology highlighting individual well-formed glands characteristic of Gleason pattern 3 carcinoma • microscopic view of a prostate tissue with discrete uniformly sized and shaped glands consistent with Gleason pattern 3 prostate cancer • focus on Gleason pattern 3 prostate cancer showing relatively distinct glandular structures with minimal variability • HE section of prostate tissue demonstrating areas of Gleason pattern 3 adenocarcinoma marked by separate infiltrating glands
	3	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 4 prostate cancer • small region of a prostate on histopathology highlighting areas of fused glands, poorly formed glands, or cribriform architecture consistent with Gleason pattern 4 carcinoma • microscopic view of prostate showcasing irregular glandular structures with varying shapes and sizes indicative of Gleason pattern 4 prostate cancer • focus on Gleason pattern 4 prostate cancer demonstrating loss of normal glandular architecture and a more disorganized growth pattern • HE section of prostate tissue demonstrating regions of Gleason pattern 4 adenocarcinoma characterized by fused or poorly defined glands

(Continued on next page)

(Continued)

Condition	Class	Prompt
	4	<ul style="list-style-type: none"> • HE histopathology image demonstrating Gleason pattern 5 prostate cancer • small region of prostate on histopathology highlighting presence of sheets of tumor cells lacking any glandular formation characteristic of Gleason pattern 5 carcinoma • microscopic view of prostate showing solid areas of tumor cells with minimal cellular pleomorphism representing Gleason pattern 5 prostate cancer • focus on Gleason pattern 5 prostate cancer demonstrating poorly differentiated cells and absence of gland formation • HE section of prostate with regions of Gleason pattern 5 adenocarcinoma characterized by a complete lack of glandular differentiation and presence of solid tumor cell clusters

The Cancer Genome Atlas (TCGA) Study Types

	1	<ul style="list-style-type: none"> • microscopic view of bladder specimen demonstrating bladder cancer • microscopic view of bladder specimen demonstrating urothelial cell carcinoma • microscopic view of bladder specimen demonstrating urothelial cancer • microscopic view of bladder specimen demonstrating urinary bladder tumor • microscopic view of bladder specimen demonstrating transitional cell carcinoma
	2	<ul style="list-style-type: none"> • microscopic view of breast specimen demonstrating breast cancer • microscopic view of breast specimen demonstrating invasive breast carcinoma • microscopic view of breast specimen demonstrating carcinoma of the breast
	3	<ul style="list-style-type: none"> • microscopic view of colon specimen demonstrating colon cancer • microscopic view of colon specimen demonstrating colorectal cancer • microscopic view of colon specimen demonstrating colon adenocarcinoma

(Continued on next page)

(Continued)

Condition	Class	Prompt
	4	<ul style="list-style-type: none"> • microscopic view of head and neck squamous cell carcinoma
	5	<ul style="list-style-type: none"> • microscopic view of kidney specimen demonstrating kidney renal clear cell carcinoma • microscopic view of kidney specimen demonstrating clear cell carcinoma • microscopic view of kidney specimen demonstrating renal clear cell carcinoma • microscopic view of kidney specimen demonstrating renal cell carcinoma clear cell type • microscopic view of kidney specimen demonstrating RCC • microscopic view of kidney specimen demonstrating kidney cancer
	6	<ul style="list-style-type: none"> • microscopic view of liver specimen demonstrating liver hepatocellular carcinoma • microscopic view of liver specimen demonstrating liver cancer • microscopic view of liver specimen demonstrating liver carcinoma • microscopic view of liver specimen demonstrating hepatocellular carcinoma
	7	<ul style="list-style-type: none"> • microscopic view of lung specimen demonstrating lung adenocarcinoma • microscopic view of lung specimen demonstrating NSCLC adenocarcinoma • microscopic view of lung specimen demonstrating lung adenocarcinoma
	8	<ul style="list-style-type: none"> • microscopic view of lung specimen demonstrating lung squamous cell carcinoma • microscopic view of lung specimen demonstrating NSCLC squamous carcinoma • microscopic view of lung specimen demonstrating squamous cell lung cancer

(Continued on next page)

(Continued)

Condition	Class	Prompt
	9	<ul style="list-style-type: none"> • microscopic view of ovary specimen demonstrating ovarian cancer • microscopic view of ovary specimen demonstrating ovarian serous cancer • microscopic view of ovary specimen demonstrating serous carcinoma • microscopic view of ovary specimen demonstrating ovarian serous cystadenocarcinoma
	10	<ul style="list-style-type: none"> • microscopic view of stomach specimen demonstrating stomach adenocarcinoma • microscopic view of stomach specimen demonstrating stomach cancer • microscopic view of stomach specimen demonstrating gastric cancer • microscopic view of stomach specimen demonstrating gastric carcinoma

Tissue Types

	1	<ul style="list-style-type: none"> • HE histopathology image from the appendix • microscopic region of a cross-section of the appendix • microscopic view of the appendix • portion of the appendix on histopathology • HE-stained image of the appendix showing the tubular structure
	2	<ul style="list-style-type: none"> • HE histopathology image of breast tissue • microscopic view of breast tissue • region of breast tissue on histopathology • HE section of breast tissue
	3	<ul style="list-style-type: none"> • HE histopathology image of the cervix • microscopic view of the cervix demonstrating epithelial cells • cervical tissue on histopathology • region of HE-stained section of the cervix

(Continued on next page)

(Continued)

Condition	Class	Prompt
	4	<ul style="list-style-type: none"> • HE histopathology image of the colon • colonic mucosa region on histopathology • microscopic view of the colon • HE-stained image highlighting portion of colon tissue • colon tissue on histopathology
	5	<ul style="list-style-type: none"> • HE histopathology image of the fallopian tube • microscopic cross-section of the fallopian tube • microscopic view of the fallopian tube • HE-stained section of the fallopian tube • fallopian tube tissue on histopathology
	6	<ul style="list-style-type: none"> • HE histopathology image of the gallbladder • microscopic view of the gallbladder • HE-stained section showing a region of the gallbladder • gallbladder tissue on histopathology
	7	<ul style="list-style-type: none"> • HE histopathology image of the liver • microscopic view of liver tissue demonstrating hepatocytes arranged in cords • liver tissue on histopathology
	8	<ul style="list-style-type: none"> • HE histopathology image of a lymph node • microscopic view of a lymph node • lymph node demonstrating mixture of immune cells predominantly lymphocytes • HE-stained section of a lymph node • lymph node tissue on histopathology
	9	<ul style="list-style-type: none"> • HE histopathology image of the ovary • microscopic view of the ovary • microscopic region of ovarian tissue • HE-stained section of the ovary • ovarian tissue on histopathology

(Continued on next page)

(Continued)

Condition	Class	Prompt
	10	<ul style="list-style-type: none"> • HE histopathology image of the placenta • placental tissue showcasing chorionic villi • microscopic view of the placenta • HE-stained section of the placenta • placental tissue on histopathology
	11	<ul style="list-style-type: none"> • HE histopathology image of the prostate • microscopic view of prostate tissue • small region of prostate tissue with acini and stroma • HE-stained section of the prostate • prostate tissue on histopathology
	12	<ul style="list-style-type: none"> • HE histopathology image of the skin • microscopic view of the skin • HE-stained section of the skin • skin tissue on histopathology
	13	<ul style="list-style-type: none"> • HE histopathology image of the thyroid • microscopic view of the thyroid • HE-stained section of the thyroid • thyroid tissue on histopathology
	14	<ul style="list-style-type: none"> • HE histopathology image of the upper GI tract such as the stomach • upper GI mucosa region on histopathology • microscopic view of the upper GI tract • HE-stained image highlighting portion of upper GI tract tissue • upper GI tract tissue on histopathology • stomach or small intestine tissue on histopathology
	15	<ul style="list-style-type: none"> • HE histopathology image of the uterus • microscopic view of the cells of the uterus • HE-stained section of the uterus • uterine tissue on histopathology
	16	<ul style="list-style-type: none"> • HE histopathology image of the vas deferens • microscopic view of the vas deferens • HE-stained section of the vas deferens • vas deferens tissue on histopathology

Table A6 | Zero shot prompts for text-based medical evaluations.

Dataset	Prompt
MedQA	"Instructions: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: " + <QUESTION> + " Response (think step by step and then end with "Final Answer:" followed by *only* the letter corresponding to the correct answer enclosed in parentheses)"
PubMedQA	"Instructions: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Answer the following question given the context (reply with one of the options): Context: " + <CONTEXT> " + Question: " + <QUESTION> + " Response (think step by step and then end with "Final Answer:" followed by *only* the letter corresponding to the correct answer enclosed in parentheses)"
MedMCQA	"Instructions: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: " + <QUESTION> + "Response (think step by step and then end with "Final Answer:" followed by *only* the letter corresponding to the correct answer enclosed in parentheses)"
MMLU	"Instructions: The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: " + <QUESTION> + " Response (think step by step and then end with "Final Answer:" followed by *only* the letter corresponding to the correct answer enclosed in parentheses)"
MedXpertQA	<QUESTION> + " Response (think step by step through each of the multiple choice options. You MUST end your response with 'Final Answer:' followed by *only* the letter corresponding to the correct answer enclosed in parentheses) like 'Final Answer:(X)'."

Table A7 | Zero-shot prompts for image classification and VQA evaluations. For all but the medical MedGemma and Med-Gemini models, the prompts below were prefixed with the following system message (SM): MI: "You are a helpful medical assistant.", or RI: "You are a helpful radiology assistant.", as indicated.

Dataset(s)	SM	Prompt/Example
MIMIC-CXR, CheXpert	MI	<p><IMAGE> + " Is there <CONDITION> in this image? You may write out your argument before stating your final very short, definitive, and concise answer (if possible, a single word or the letter corresponding to your answer choice) X in the format "Final Answer: X":</p> <p>Where <CONDITION> is one of the following:</p> <ul style="list-style-type: none"> • For MIMIC-CXR & CheXpert: atelectasis, cardiomegaly, consolidation, edema, pleural effusion. • For CXR14: lung opacity, pneumothorax. <p><i>Exception for CXR14:</i> The fracture condition uses the question "Are there any fractures in the image?".</p>
PathMCQA	MI	<p>"<QUESTION>"</p> <p>Example for <QUESTION>: "What is the predominant diagnostic finding in this cervical biopsy image? Please select the correct option from the following choices: A) Normal or benign tissue B) Mild dysplasia (CIN1) C) Moderate or severe dysplasia (CIN 2+)"</p>
US-Derm MCQA	MI	<p><IMAGE> + "<QUESTION> You may write out your argument before deciding on the most likely condition X (with X being one of: A, B, C, or D) in the format: "The most likely diagnosis is: X":</p> <p>Example for <QUESTION>: "Question: Given the following image, what is the most likely dermatology condition? Options: (A) Granuloma annulare. (B) Melanocytic Nevus. (C) Erythema annulare centrifugum. (D) Morphea/Scleroderma."</p>
EyePACS	MI	<p><IMAGE> + "Given this fundus image, determine the most likely diabetic retinopathy (DR) stage present, even if you are unsure:</p> <p>A: No DR B: mild DR C: moderate DR D: severe DR E: proliferative DR</p> <p>You may write out your argument before stating your final, very short, definitive, and concise answer (no more than a few words) and letter corresponding to your answer choice X in the format "The most likely diagnosis is: X":</p>
SLAKE	MI	<p><IMAGE> + <QUESTION> + "You may write out your argument before stating your final very short, definitive, and concise answer (if possible, a single word or the letter corresponding to your answer choice) X in the format "Final Answer: X": "</p>
VQA-Rad	RI	<p><IMAGE> + "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, definitive, and concise answer (if possible, a single word) to the following question: <QUESTION>"</p>
MedXpertQA (multi-modal questions)	MI	<p>"Figure A" + <IMAGE_A> + "Figure B" + <IMAGE_B> + ... + "<QUESTION> Think step by step through each of the multiple choice options. You MUST end your response with 'Final Answer:' followed by only the letter corresponding to the correct answer enclosed in parentheses)."</p>
CXR report generation	-	<p><IMAGE> "<INDICATION> findings:"</p> <p>Where <INDICATION> is the indication section of the report as prefix.</p>

C. Comparison of CXR data-efficient learning

To compare with previous research on CXR image encoder-based data-efficient learning, we compared MedSigLIP to ELIXR (Xu et al., 2023). Performance of MedSigLIP and ELIXR are compared for various CXR findings at different training set sizes on CheXpert (Figure A1) and CXR14 (Figure A2).

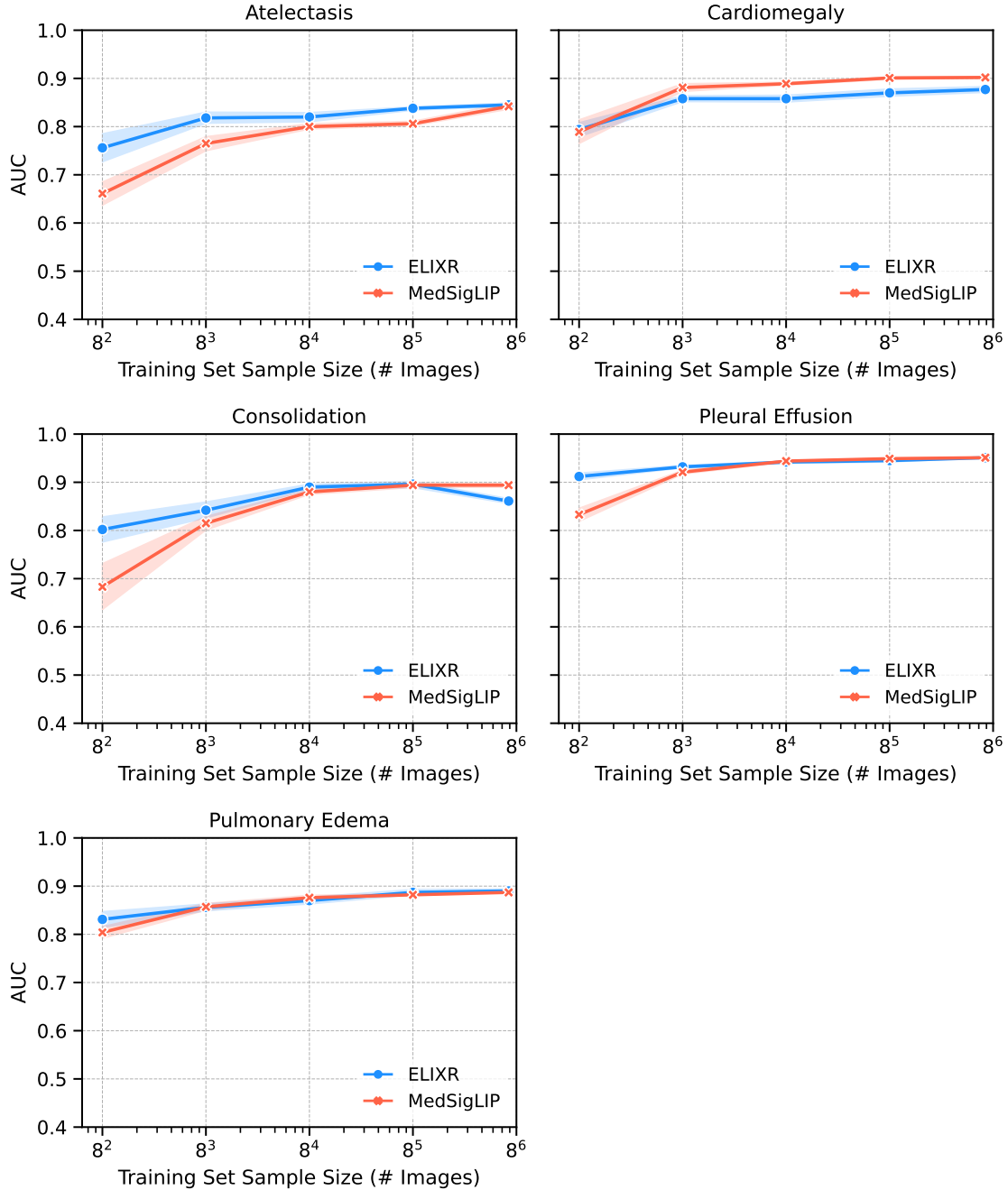


Figure A1 | Individual results for CXR data efficient learning on CheXpert datasets comparing to ELIXR (Xu et al., 2023)

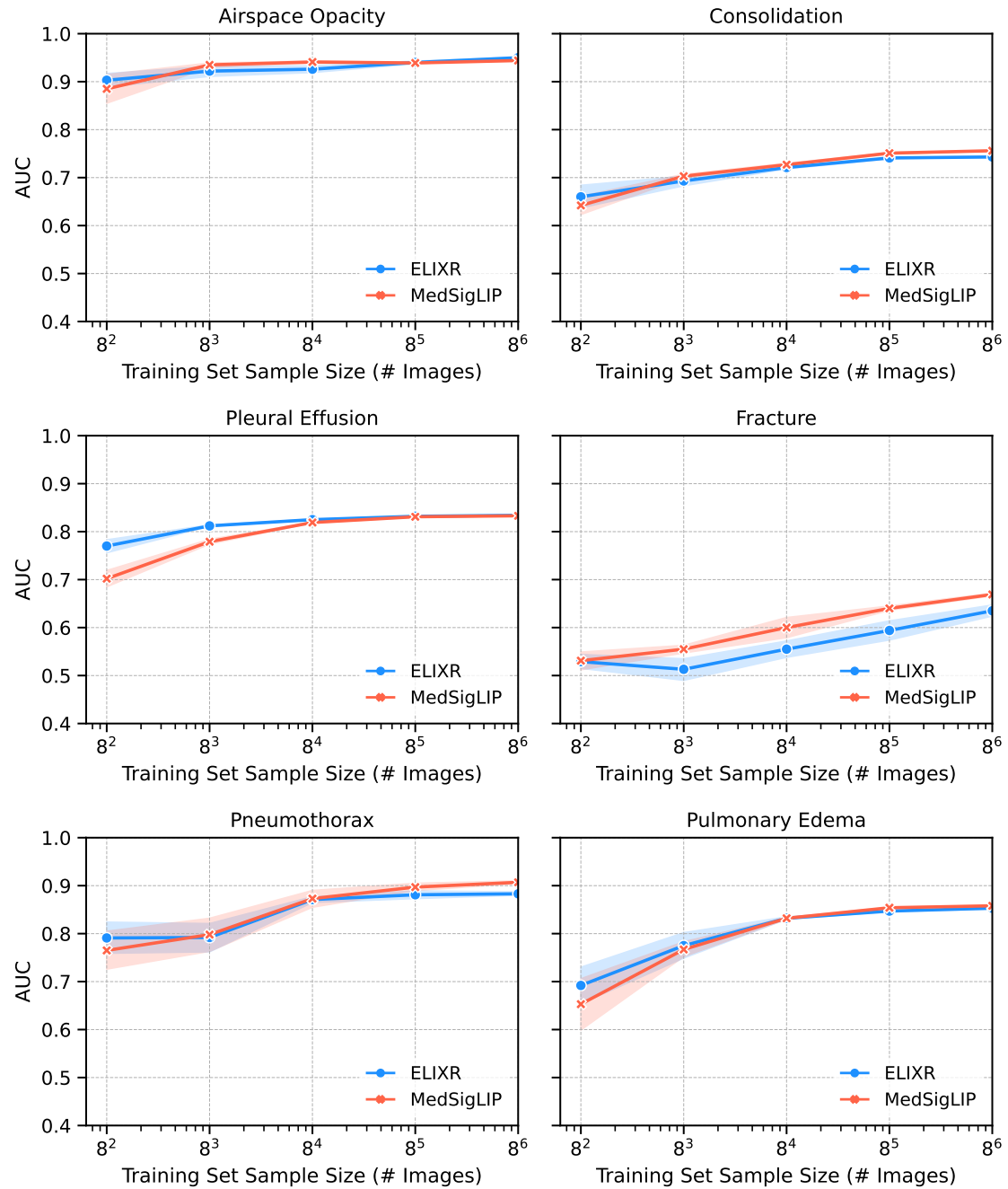


Figure A2 | Individual results for CXR data efficient learning on CXR14 datasets comparing to ELIXR (Xu et al., 2023)

D. Additional details about EHRQA

EHRQA detailed results appear in Figure A3, where the base MedGemma 27B text model is compared against the fine-tuned MedGemma 27B model and Gemini 2.5 Pro. Details of EHRQA question categories are shown in Table A8 and prompts for question rewrites and evaluation are shown in Table A9.

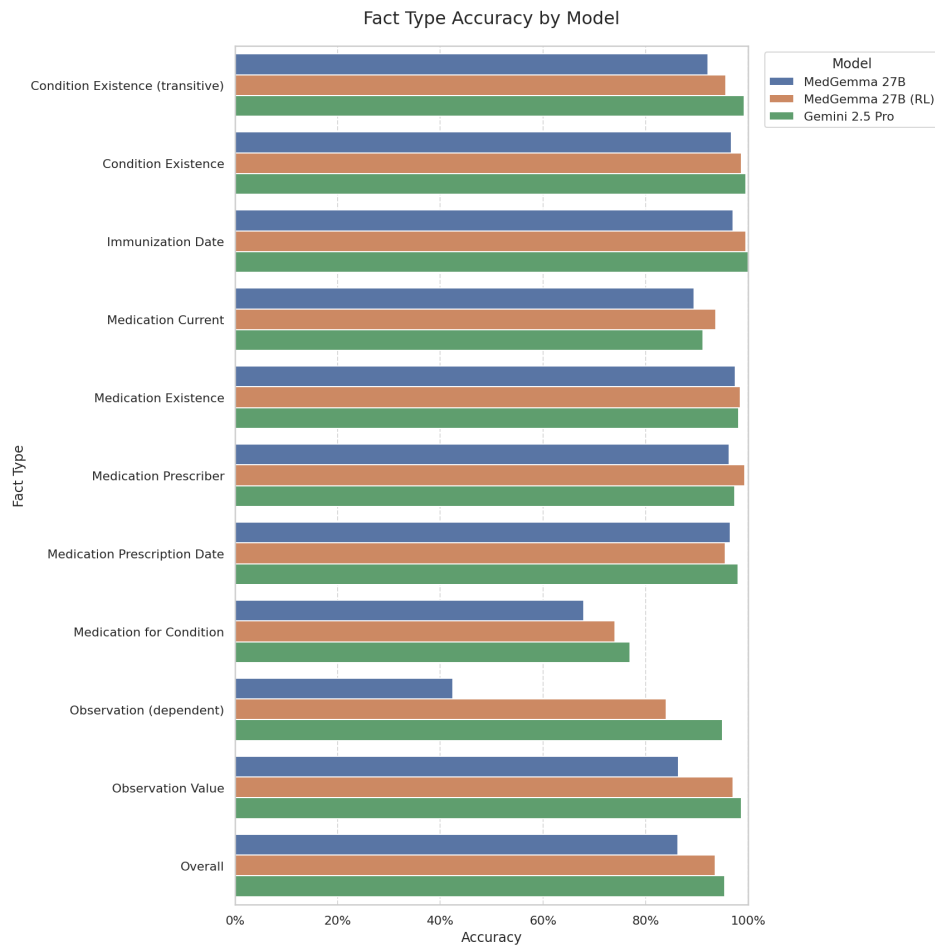


Figure A3 | EHRQA detailed results, comparing MedGemma 27B (text-only) before and after RL-tuning, along with Gemini 2.5 Pro

Table A8 | EHRQA Question Categories with Original (Template-Generated) and LLM-Rephrased Examples.

Category	Description	Example: Templated	Example: LLM rephrased
Condition existence (transitive)	Questions related to whether a patient has a history of a general class of medical conditions.	Has the patient had a medical condition in gingival structure in the past?	Has the patient had a past medical condition affecting the gums?
Condition existence	Questions related to whether a patient has a history of a given medical condition.	Has the patient had acute bronchitis in the past?	Has the patient had acute bronchitis previously?
Immunization date	Questions related to the existence of patient immunizations and their immunization dates.	When did the patient last receive Hep B, adult immunization?	When was the patient's last adult hepatitis B vaccination administered?
Medication current	Questions related to whether the patient has active prescriptions for a given medication.	Does the patient have a prescription for {28 (norethindrone 0.35 MG Oral Tablet) } Pack?	Does the patient have an active prescription for norethindrone oral tablet?
Medication existence	Questions related to whether the patient has ever taken a given medication or type of medication.	Did the patient have a prescription for insulin isophane, human 70 UNT/ML / insulin, regular, human 30 UNT/ML Injectable Suspension in the past?	Did the patient have a prescription for human insulin isophane / human insulin regular injectable suspension in the past?
Medication prescriber	Questions related to the prescriber of a given medication for a patient.	Who prescribed acetaminophen 325 MG Oral Tablet on the date 2022-06-30?	Who prescribed acetaminophen on June 30, 2022?
Medication prescription date	Questions related to the medication prescription dates.	When was acetaminophen 325 MG Oral Tablet first prescribed?	When was acetaminophen oral tablet first prescribed?
Medication for condition	Questions on the relationships between medication and medical condition history.	Which condition were the medications acetaminophen 300 MG / hydrocodone bitartrate 5 MG Oral Tablet prescribed for?	For what condition was acetaminophen/hydrocodone prescribed?
Observation (dependent)	Questions related to patient lab observations related to some medical event.	What was the patient's first recorded value for Pain severity - 0-10 verbal numeric rating [Score] - Reported after they were first prescribed Acetaminophen 325 MG Oral Tablet?	What was the patient's initial reported pain score after the first prescription of acetaminophen 325 mg oral tablet?
Observation value	Questions related to patient lab observations and their values.	What is the patient's most recent recorded value for Heart rate?	What is the patient's most recent heart rate reading?

Table A9 | Prompts for EHRQA construction and evaluation.

Purpose	Prompt
General question rewrite	<p>You are a helpful medical editing assistant.</p> <p>You will be presented with a question which pertains to the medical record for a patient. Rewrite the question in a way that sounds more natural, omitting unnecessary (in particular parenthesized) details but retaining enough detail to leave the question unambiguous.</p> <ul style="list-style-type: none"> - Use correct capitalization and language appropriate for a medical professional. - Avoid contractions in the question, but use common-language terminology where that is possible without causing ambiguity. - Keep in mind that the question may be a multiple choice question followed by several answer choices. - Be sure the rewritten questions ask about the same topic as the original. For instance, if there is a question about who prescribed a medication, the re-written question should do the same. <p>ORIGINAL QUESTION: <ORIGINAL_QUESTION> REWRITTEN QUESTION:</p>
Medication question rewrite	<p>You are a helpful medical editing assistant.</p> <p>You will be presented with a question which pertains to the medical record for a patient. Rewrite the question in a way that sounds more natural, omitting unnecessary (in particular parenthesized) details but retaining enough detail to leave the question unambiguous.</p> <ul style="list-style-type: none"> - Use correct capitalization and language appropriate for a medical professional. - Avoid contractions in the question, but use common-language terminology where that is possible without causing ambiguity. - Keep in mind that the question may be a multiple choice question followed by several answer choices. - If the brand name is mentioned in the original question, rewrite the question using that brand name only. - Remove the medication strength/dosage information in the medication name. - Use common names for medication classes if available. <p>EXAMPLES: <FEW_SHOT_EXAMPLES> ORIGINAL QUESTION: <ORIGINAL_QUESTION> REWRITTEN QUESTION:</p>
Evaluation	<p>You are a medical assistant specializing in answer questions about a patient electronic health records (EHR). You will be provided with a patient's full EHR data and a question for you to answer.</p> <p>===== Patient EHR Data ===== <PATIENT_CONTEXT> ===== End Patient EHR Data =====</p> <p>Based on the data provided in the patient's EHR, please answer concisely with no explanation. If the question is a multiple choice question, answer with only the letter of the correct choice, such as "B". If the question is a free response question, answer with the correct data. In either case, remember to not provide an explanation.</p> <p>Question: <QUESTION> Answer choices: ANSWER_CHOICES</p>

E. Additional medical reasoning examples

Tables A10, A11, A12 and A13 show additional medical reasoning examples, with clinical commentary, in the same form as Tables 5 and 6.

Table A10 | Example of MedGemma 27B open clinical reasoning in a cardiology setting, including commentary by a US board-certified practicing cardiologist

Prompt

A 55-year old woman with active smoking history, type II diabetes, hypertension, history of coronary artery disease was recently hospitalized for heart failure with an ejection fraction of 35%. She is presenting for follow-up post discharge. Her discharge medications included lisinopril 10 mg daily, carvedilol 2.5 mg twice daily, and eplerenone 25 mg daily, in addition to her usual medications for diabetes. In clinic, she reports feeling symptomatic with dyspnea after going up 1 flight of stairs though is asymptomatic at rest. Vital signs at rest: BP 130/90, HR 75, RR 10, Temperature 97.2F, SpO2 98% on room air. Labs are remarkable for Na 132 mEq/L, K 4.9 mEq/L, BUN 22 mg/dL, Cr 1.04 mg/dL. The patient reports having a friend on Entresto. Please review whether Entresto is indicated in this patient and why or why not.

Continued on next page

Table A10 – Continued from previous page

MedGemma 27B output

Okay, let's review the indication for Entresto (sacubitril/valsartan) in this patient.

Patient Summary:

- **Diagnosis:** Heart Failure with Reduced Ejection Fraction (HFrEF), EF 35%.
- **Symptoms:** Dyspnea on exertion (1 flight of stairs), NYHA Class II-III.
- **Current HFrEF Medications:** Lisinopril (ACE inhibitor), Carvedilol (Beta-blocker), Eplerenone (MRA).
- **Comorbidities:** Smoking, Type II Diabetes, Hypertension, CAD.
- **Relevant Labs:** K 4.9 mEq/L (high-normal), Cr 1.04 mg/dL (mildly elevated/high-normal).
- **Vitals:** BP 130/90 mmHg (acceptable for initiation).

Entresto Indication Review:

Entresto (sacubitril/valsartan) is indicated for the reduction of cardiovascular death and hospitalization for heart failure in adult patients with chronic heart failure (NYHA class II-IV) with reduced ejection fraction (LVEF \leq 40%).

Analysis for this Patient:

1. **Ejection Fraction:** The patient's EF is 35%, which meets the criteria (\leq 40%).
2. **Symptom Status:** The patient is symptomatic (dyspnea on exertion) despite being on foundational guideline-directed medical therapy (GDMT) including an ACE inhibitor (lisinopril), a beta-blocker (carvedilol), and an MRA (eplerenone). This is the target population for Entresto.
3. **Contraindications:**
 - **ACE Inhibitor Use:** Entresto should *not* be used concurrently with an ACE inhibitor (like lisinopril) due to increased risk of angioedema, hypotension, and hyperkalemia. The lisinopril would need to be stopped, typically with a 36-hour washout period, before starting Entresto.
 - **Angioedema History:** No history of angioedema related to ACE inhibitors or ARBs is mentioned.
 - **Hypotension:** Her current BP of 130/90 mmHg is generally acceptable for initiating Entresto, although close monitoring is needed.
 - **Renal Function:** Her creatinine is 1.04 mg/dL. While Entresto can cause a small increase in creatinine, this level is not a contraindication, but renal function should be monitored.
 - **Potassium:** Her potassium is 4.9 mEq/L (high-normal). This requires close monitoring, as Entresto can increase potassium levels, especially in combination with an MRA like eplerenone.

Conclusion:

Yes, Entresto is indicated for this patient.

Reasoning: She has symptomatic HFrEF (EF 35%) despite being on appropriate foundational therapy (ACE inhibitor, beta-blocker, MRA). Clinical trials have demonstrated that switching from an ACE inhibitor (like lisinopril) to Entresto significantly reduces morbidity and mortality in this patient population.

Important Considerations:

- **Lisinopril Discontinuation:** The lisinopril must be stopped before starting Entresto.
- **Monitoring:** Close monitoring of blood pressure, heart rate, renal function (creatinine/eGFR), and potassium levels is crucial, especially after initiation and dose titration.
- **Dosing:** Entresto is typically started at a low dose (e.g., 24/25 mg BID) and titrated upwards to the target dose (e.g., 49/51 mg BID or 97/103 mg BID) as tolerated, based on guideline recommendations.

This represents a significant therapeutic opportunity to improve her heart failure outcomes.

Commentary by a US board-certified practicing cardiologist

Strengths: Correctly states that entresto is indicated for the reduction of CV death and HF hospitalization. Appropriately checks renal function, electrolytes, angioedema and BP. Lists washout and monitoring.

Areas for Improvement: Dose is incorrect for the lowest dose of entresto. Minor nit that entresto has now been shown in meta-analysis to reduce all cause mortality in EF < 40% (not just cardiac mortality as stated) (Evdavaykha et al., 2025). Creatinine is mentioned but estimated GFR is the more useful measure. Minor but no mention of carvedilol dose. The dose is very low and should be increased. If it is increased, there will be less BP room for entresto so would need to titrate both carefully but can be done rapidly (Strong HF trial). The other thing that could have been mentioned is empagliflozin which is not included in this list. But the question is specific to entresto so understandable.

Table A11 | Example of MedGemma 27B open clinical reasoning in a cardiology setting with *succinct* output, including commentary by a US board-certified practicing cardiologist

Prompt

Same as in Table A10 with the suffix: Please be succinct.

MedGemma 27B output

Yes, Entresto is indicated.

Rationale: This patient has symptomatic Heart Failure with reduced Ejection Fraction (HFrEF) (EF 35%, NYHA Class II) despite being on foundational guideline-directed medical therapy (ACE inhibitor, beta-blocker, MRA). Entresto (sacubitril/valsartan) is recommended to replace the ACE inhibitor (lisinopril) in such patients to further reduce cardiovascular death and heart failure hospitalizations. Her current BP and potassium levels are acceptable for initiation, although close monitoring is required.

Clinical commentary by a US board-certified practicing cardiologist

Strengths: A more useful answer - short and to the point, including the rationale for the treatment.

Areas for Improvement: Does not include renal function in the list of things to monitor (just BP and potassium). Same note as above that entresto has now been shown in meta-analysis to reduce all cause mortality in EF<40% rather than just cardiac mortality as stated.

Table A12 | Example of MedGemma 27B open clinical reasoning in an infectious disease setting, including commentary by a UK Specialist Registrar in infectious diseases

Prompt

A 17-year old otherwise healthy man presents in the emergency room for fatigue and spiking fevers. Two weeks prior, he reports having felt flu-like symptoms with fever, nausea, vomiting, and cough. Symptoms persisted but have worsened, with new onset shortness of breath, headache, and abdominal discomfort. There have been no sick contacts at home. He has been backpacking across various regions in the United States, most recently in Hawaii.

On evaluation, vital signs are remarkable for SBP 115/85, HR 110, RR 14, Temperature 102.2F, SpO2 90% on room air. Exam is remarkable for bilateral crackles and icterus.

Labs show the following:

- WBC: $18.2 \times 10^3/\text{mL}$, Hb: 15.5 g/dL, Plt: $120 \times 10^3/\text{mL}$
- Na: 142 mEq/L, K: 5.1 mEq/L, Cl: 101 mEq/L, CO₂: 26 mEq/L, BUN: 46 mg/dL, Cr: 2.2 mg/dL
- ALP: 84 U/L, ALT: 175 U/L, AST: 140 U/L, Total bilirubin: 2.0 mg/dL

CXR shows bilateral pulmonary infiltrates. A follow-up echocardiogram shows EF 40%. A respiratory viral panel was negative. Blood, sputum, and urine cultures are taken.

What would be the next best steps in terms of diagnostic testing to establish the etiology of his symptoms?

Continued on next page

Table A12 – Continued from previous page

MedGemma 27B output**Rationale Breakdown:**

1. **Clinical Presentation Analysis:** The patient presents with a severe, multi-system illness characterized by fever, fatigue, shortness of breath, headache, abdominal discomfort, jaundice (icterus), hypoxia, leukocytosis, thrombocytopenia, acute kidney injury (elevated BUN/Cr), hepatitis (elevated ALT/AST/bilirubin), and possible myocarditis (reduced EF on echo). The illness began with flu-like symptoms two weeks ago and has progressively worsened.
2. **Key Clues:**
 - **Travel History:** Recent travel to Hawaii is a significant clue, pointing towards infections endemic to that region.
 - **Multi-organ Involvement:** The combination of pulmonary infiltrates, jaundice, AKI, and myocarditis suggests a systemic process, often infectious.
 - **Specific Lab Findings:** Leukocytosis with thrombocytopenia, elevated liver enzymes, and bilirubin are important diagnostic pointers.
3. **Differential Diagnosis (Focusing on Hawaii-Endemic Infections):**
 - **Leptospirosis:** Highly consistent with the presentation. Causes fever, headache, myalgia, conjunctival suffusion (not mentioned but common), jaundice, renal failure, pulmonary hemorrhage/ARDS, and myocarditis. Thrombocytopenia is common; leukocytosis can occur. Hawaii is an endemic area.
 - **Dengue Fever:** Endemic in Hawaii. Causes fever, headache, myalgia, rash (not mentioned), leukopenia (patient has leukocytosis), thrombocytopenia. Severe dengue can cause organ impairment, but the specific combination of jaundice, AKI, and myocarditis with leukocytosis makes it less likely than leptospirosis.
 - **Hantavirus Pulmonary Syndrome (HPS):** Endemic in Hawaii. Causes fever, myalgia, headache, followed by respiratory failure and pulmonary edema. Thrombocytopenia and leukocytosis are typical. However, significant jaundice and myocarditis are less characteristic of HPS.
 - **Rickettsial Infections (e.g., Murine Typhus, Scrub Typhus):** Endemic in Hawaii. Can cause fever, headache, rash, myalgia, sometimes pneumonia, hepatitis, AKI.
 - **Other possibilities:** Influenza (severe complications), other viral hepatitis (EBV, CMV), bacterial sepsis, melioidosis (rare but reported).
4. **Prioritizing Diagnostic Tests:** Based on the clinical picture and travel history, Leptospirosis is the most likely diagnosis. Therefore, specific testing for Leptospirosis should be prioritized. Dengue and Hantavirus are also important considerations due to endemicity and overlapping features.
5. **Specific Diagnostic Tests:**
 - **Leptospirosis:**
 - **Serology:** ELISA for IgM antibodies is often the first step for faster results. Microscopic Agglutination Test (MAT) is the gold standard but takes longer. Paired sera (acute and convalescent) are often needed for definitive diagnosis.
 - **PCR:** Detects *Leptospira* DNA in blood (early) or urine (can be positive later, especially after the first week). Given the patient is 2 weeks into symptoms, urine PCR might be useful.
 - **Dengue:**
 - **Serology:** IgM/IgG antibodies.
 - **NS1 Antigen:** Detectable early in illness.
 - **PCR:** Detects viral RNA early in illness.
 - **Hantavirus:**
 - **Serology:** IgM/IgG antibodies.
 - **PCR:** Detects viral RNA early in illness.
 - **Other Tests:** While awaiting results for the primary suspects, continue monitoring organ function (liver, kidney), coagulation, and review pending cultures. Consider additional tests like EBV/CMV serology, HIV testing, and Rickettsial serology/PCR if the initial workup is negative or the clinical picture evolves.
6. **Conclusion:** The most crucial next steps are to order specific diagnostic tests for the most likely endemic infections given the patient's presentation and travel history, particularly Leptospirosis, Dengue, and Hantavirus.

Final Answer: The final answer is

Order specific serological and PCR testing for Leptospirosis, Dengue, and Hantavirus

Continued on next page

Table A12 – Continued from previous page

Commentary by a UK Specialist Registrar in infectious diseases

Strengths: Well-structured answer, making the information easy to parse. Correctly has the most likely diagnosis (lepto) as the focus of the answer. The suggested diagnostic testing enables confirmation of the specific diagnoses listed in the differential.

Areas for Improvement: Dengue fever, hantavirus and melioidosis don't seem to occur in Hawaii much, if at all (and at least, they are not endemic, as stated in the answer). The suggested diagnostic testing perhaps narrows towards specific diagnoses too quickly, whereas in reality broader testing may still be necessary (e.g. CSF analysis given headache, further chest imaging via CT, etc.) alongside more specific confirmatory tests. It could mention that many of these diagnostic tests are relatively unusual in most hospitals, often requiring samples to be sent via courier to specialized labs, and therefore there is a significant time lag associated with getting results. This may necessitate empiric treatment whilst awaiting for results, which is an important clinical consideration.

Table A13 | Example of MedGemma 27B open clinical reasoning in an infectious disease setting with *succinct* output, including commentary by a UK Specialist Registrar in infectious diseases

Prompt

Same as in Table A12 with the suffix: Please be succinct.

MedGemma 27B output**Next best steps in diagnostic testing:**

1. Leptospirosis serology (IgM ELISA/MAT).
2. Hantavirus serology (IgM).
3. Dengue serology (IgM/IgG, NS1 antigen).
4. Blood and urine cultures.
5. Urinalysis.
6. Cardiac enzymes (Troponin).
7. Coagulation studies.

Commentary by a UK Specialist Registrar in infectious diseases

Strengths: Correctly lists confirmatory diagnostic tests for each suggested differential. Covers both specific confirmatory tests (to pinpoint a particular condition) and more generalized diagnostics (e.g. coagulation studies) that would yield more useful information to guide next steps in case the confirmatory diagnostic tests are negative.

Areas for Improvement: Same feedback as before regarding the epidemiology of hantavirus/dengue in Hawaii - seems to be rare, if at all present. Repeats several items already listed as being performed in the query (e.g. blood, urine cultures). Unclear why it limits the infectious disease testing to a single antibody (IgM) for each potential cause - it would be normal to perform IgM and IgG as both are useful in distinguishing between an acute infection versus a prior infection. No mention of molecular (genetic) testing which is increasingly used for diagnosis.

F. MedGemma 27B multimodal model

In addition to the MedGemma 4B multimodal and 27B text-only variants described earlier in this report, we are also releasing a MedGemma 27B multimodal variant. The training methodology for this variant was the same as for the MedGemma 4B multimodal model with the addition of two training datasets: EHRQA (details in Section 5.2), to improve the model’s inherent EHR understanding, and Chest ImaGenome (Goldberger et al., 2000; Wu et al., 2021), to enable anatomy localization on chest X-ray images.

The MedGemma 27B multimodal variant provides all of the capabilities of the 4B multimodal variant but with significantly improved language capabilities as well as improved EHR understanding and anatomy localization. Formal evaluation of this variant is ongoing but preliminary results can be found in Table A14.

Table A14 | Preliminary results for MedGemma 27B multimodal model

Task	Metric	MedGemma 4B Multimodal	MedGemma 27B Text-only	MedGemma 27B Multimodal
Text evaluation				
MedQA (4-op)	Accuracy	64.4	87.7	85.3
MedMCQA	Accuracy	55.7	74.2	70.2
PubMedQA	Accuracy	73.4	76.8	77.2
MMLU Med	Accuracy	70.0	87.0	86.2
MedXpertQA (text only)	Accuracy	14.2	25.7	23.7
AfriMed-QA	Accuracy	52.0	84.0	72.0
Electronic health record information retrieval				
EHRQA	Accuracy	67.6	86.3	90.5
Medical image classification				
MIMIC CXR	Average F1 for 5 conditions	88.9	N/A	90.0
CheXpert CXR	Average F1 for 5 conditions	48.1	N/A	49.9
CXR14	Average F1 for 3 conditions	50.1	N/A	45.3
DermMCQA	Accuracy	71.8	N/A	71.7
PathMCQA	Accuracy	69.8	N/A	71.6
Eyepacs	Accuracy	64.9	N/A	75.3
Visual question answering				
SlakeVQA	Tokenized F1	72.3	N/A	70.3
VQA-Rad	Tokenized F1	49.9	N/A	46.7
Knowledge and reasoning				
MedXpertQA (text + MM)	Accuracy	18.8	N/A	26.8
Report generation				
MIMIC CXR	Radgraph F1	29.5 [†]	N/A	27.0 [†]

[†] Results from the pretrained checkpoints.