

An open vision-language model for diverse medical applications

Received: 9 March 2026

Accepted: 4 August 2026

Published online: 6 October 2026

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Artificial intelligence has high potential for impact in healthcare applications, but its training and deployment are challenging due to diverse data, a complex spectrum of possible tasks and important privacy needs. High-performing foundation models that enable data-efficient fine-tuning for diverse downstream tasks can meaningfully accelerate development in this domain. We introduce MedGemma, a collection of medical vision-language foundation models based on Gemma 3. MedGemma demonstrates advanced medical understanding and reasoning across images and text and multiple medical imaging domains, exceeding the performance of similarly sized generative models while maintaining the general capabilities of the Gemma base models. For out-of-distribution tasks, MedGemma achieves improvements of 2.6–10% in medical image question answering, 15.5–18.1% in chest X-ray finding classification and 10.8% in agentic evaluations compared with the base models. Our results show that fine-tuning MedGemma can be more effective than fine-tuning the base Gemma 3 model for medical tasks, particularly in the setting of limited training data. We additionally introduce MedSigLIP, a medically tuned vision encoder derived from SigLIP. MedSigLIP powers the visual understanding capabilities of MedGemma and, as an encoder, achieves performance comparable to or better than that of many specialized medical image encoders. Taken together, the MedGemma collection provides a strong foundation of medical image and text capabilities, with the potential to accelerate medical research and the development of downstream applications.

The landscape of modern healthcare is characterized by the generation and use of unprecedented volumes and diversity of data. Diagnosis, treatment and monitoring rely on connecting the dots between many different data sources and medical specialties. Recently developed large multimodal models—advanced artificial intelligence (AI) systems trained on massive and diverse datasets—exhibit remarkable capabilities in detecting complex patterns, generating coherent text and processing visual information^{1–3}. These capabilities mark a potential paradigm shift in how AI can be used to assist with current healthcare workflows and extract new insights.

While general-purpose (nonmedically tuned) large multimodal models demonstrate impressively broad abilities, generic models can lack the nuanced medical understanding needed to interpret and reason about medical data in a robust way^{4–6}. Recognizing this

gap, we developed MedGemma, a collection of medically tuned, vision-language foundation models. Built upon the robust architecture of Gemma 3 (ref. 7), the MedGemma models are designed to interpret and reason about medical images and text while retaining the strong general-purpose capabilities of Gemma 3.

MedGemma

We developed MedGemma as a collection of three model variants well suited for a variety of practical downstream applications (Fig. 1). Specifically, these include two multimodal variants, MedGemma 4B and MedGemma 27B (4 billion and 27 billion parameter models, respectively), which can accept text, images or both as input, and MedGemma 27B Text, which is optimized for text inputs. The multimodal MedGemma variants demonstrate strong performance on

✉ e-mail: asellerg@google.com; davesteiner@google.com; dangolden@google.com; linyan@google.com

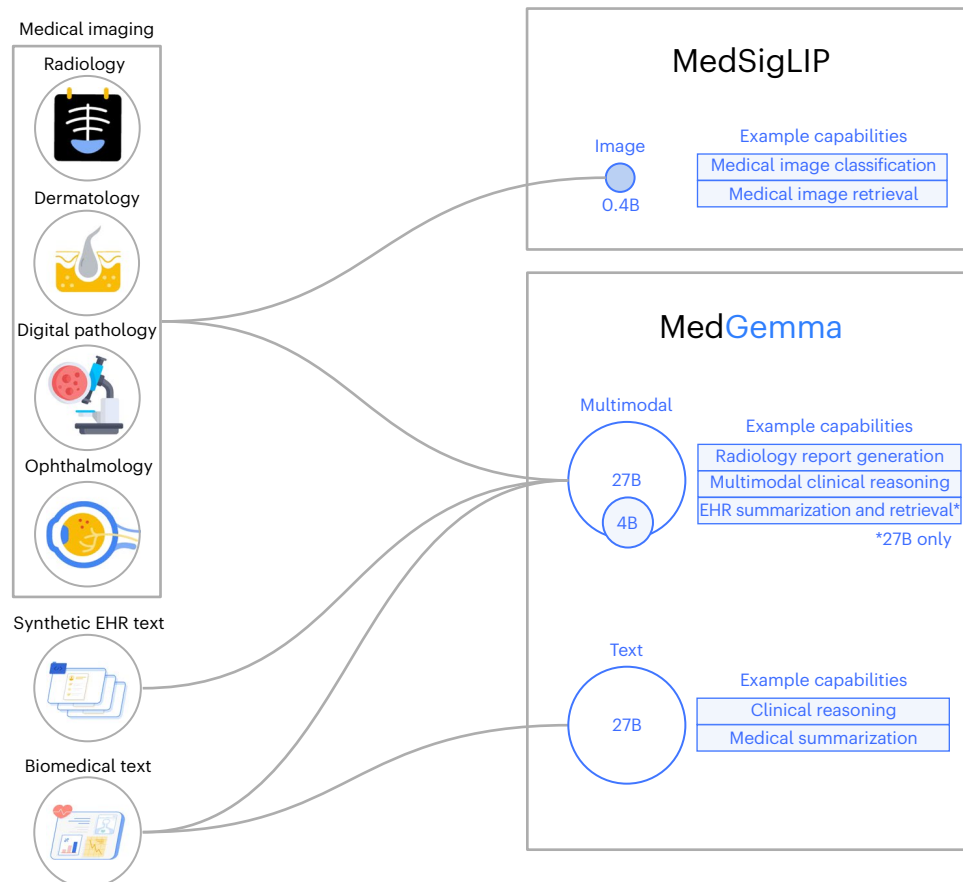


Fig. 1 | Overview of the MedGemma models. The MedGemma model collection features the MedSigLIP image encoder (400 million parameters), MedGemma 4B (4 billion parameters, multimodal), MedGemma 27B (27 billion parameters, multimodal) and MedGemma 27B Text (27 billion parameters, text only).

Modalities used for domain-specific training of the respective models are indicated by the lines connecting data types on the left to models on the right. EHR, electronic health record.

vision question answering (VQA) benchmarks, achieving performance similar to or exceeding prior models such as Med-Gemini^{8,9} and larger frontier models such as Gemini 2.5, despite being considerably smaller. MedGemma also demonstrates notable size-to-performance efficiency across benchmark tasks, including MedQA¹⁰, MedMCQA¹¹, PubMedQA¹², MMLU Med¹³, AfriMed-QA¹⁴ and AgentClinic¹⁵, outperforming open models of similar scale in our analysis.

The combination of strong benchmark results across both text and medical imaging tasks illustrates this high potential for MedGemma to be efficiently tuned and adapted to a diverse set of medical AI research and development efforts using a single base model. Additionally, MedGemma was optimized to retain the general-purpose language capabilities of Gemma 3, further enabling a particularly flexible and robust starting point for development.

MedGemma incorporates a specialized 400-million-parameter medical image encoder, MedSigLIP, specifically trained for medical imaging. To improve the encoder's capabilities for medical imaging, we fine-tuned the vision encoder in Gemma 3 (SigLIP-400M)⁷ using millions of medical image–text pairs. These training data are described in Extended Data Table 1. MedSigLIP 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 (summarized in Supplementary Note—MedSigLIP results).

Notably, these results represent only baseline capabilities, and the open nature of these models is intended to enable researchers worldwide to continue improving them for specific needs. In this regard,

we also provide illustrative examples of fine-tuning for chest X-rays (CXRs), histopathology, dermatology and electronic health record question answering.

Results

MedGemma performance across medical text and imaging benchmarks

MedGemma performance was evaluated across several medical text and medical imaging benchmarks (Table 1), including both in-distribution benchmarks, for which a subset of the data was used for training or tuning with a held-out set for evaluation, and out-of-distribution (OOD) benchmarks, for which no subset of the data was used for training or tuning (note that, for general-purpose benchmarks, it is difficult to determine whether data are OOD given the large amount of pretraining data in the base model training mixture). Overall, MedGemma achieved accuracy exceeding or on par with much larger, nonopen models as well as task-specialized models across several benchmark tasks. The sections below outline performance across different categories, including medical text question answering, medical image classification, open-ended clinical reasoning, medical visual question answering, report generation and agentic behavior. The relevant model variants are included in the result tables below, with the complete set of performance results across MedGemma variants provided in Supplementary Note—Additional training data details.

Medical text question answering. Across all text-only biomedical question-answering tasks evaluated, MedGemma demonstrated superior performance to the 'baseline' Gemma 3 model, as well as

Table 1 | Overview of MedGemma evaluation datasets

Task	Dataset	Modality	No. of examples	OOD ^a
Medical text question answering	MedQA	Text	1,273	–
	MedMCQA	Text	4,183	–
	PubMedQA	Text	500	–
	MMLU Med	Text	3,685	–
	MedXpertQA	Text	2,450	✓
	AfriMed-QA (MCQ)	Text	25	–
	EHRQA	Text	4,377	–
	MultiMedQA (subset; open ended)	Text	130	✓
Medical image classification	MIMIC-CXR (Med-Gemini test set)	Radiology	1,532	–
	MIMIC-CXR (MAIRA test set)	Radiology	2,461	–
	ChestX-ray14 (CXR14)	Radiology	1,962	✓
	CheXpert	Radiology	668	✓
	US-Derm MCQA	Dermatology	1,996	–
	PathMCQA	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	–
CXR 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

^aOOD indicates data sources not seen during any model development stages. N/A, not applicable.

competitive performance with much larger models in many cases (Table 2). Confidence intervals for the results, where calculated, are shown in Supplementary Table 1. Acknowledging that frontier models often still outperform smaller open models, these results highlight the clear benefit of domain specialization and position MedGemma as a useful starting point for many downstream applications. This evaluation was based on MedQA, MedMCQA, PubMedQA, MMLU subsets, AfriMed-QA and the OOD MedXpertQA. Notably, MedGemma's performance of 86.2 on MedQA represented the best performance for an open model at the time of its release, other than the 671-billion-parameter DeepSeek-R1 model (91.0)¹⁶, and was substantially higher than the next closest well-documented open model, OpenBioLLM 70B (78.2)¹⁷.

The open nature of MedGemma has also enabled additional third-party evaluations. For medical text question answering, the analysis provided in ref. 18 further demonstrates the performance advantages conferred by MedGemma's medical domain training relative to Gemma 3. One notable difference was the improved performance of MedGemma relative to Gemma 3 on the MedHallu benchmark evaluation¹⁹, which asks a model to distinguish between a correct answer and a feasible sounding, but incorrect answer as one way to better understand model reliability. On the established 'hard set' of the MedHallu task, MedGemma 4B scored 0.50 and Gemma 3 4B scored 0.09 (normalized performance score).

Medical image classification. Medical image interpretation is a key contribution of the multimodal MedGemma models. To provide performance insights across a diverse set of medical imaging domains, we used image-based classification tasks in radiology, histopathology, dermatology and ophthalmology. These domains were selected

because they are most consistent with training data and intended capabilities of MedGemma. (Of note, addressing classification as a zero-shot generative task may not provide maximum performance as compared to an embedding-based, task-specific classifier, but these evaluations are intended to provide insights into the quality and performance of the underlying image encoder and the overall model capabilities.) Results are summarized in Table 3 for CXR, histopathology, dermatology, and retina image classification. Notably, MedGemma not only demonstrated substantially superior performance relative to the standard Gemma 3 baselines but also outperformed much larger API-based models.

Open-ended medical question answering and clinical reasoning.

To further evaluate the quality of model output, generated responses to open-ended medical content questions were evaluated. Model responses to these questions were rated by three physicians per question (see Supplementary Note for additional details). This analysis demonstrated a relative decrease in factually inaccurate responses of approximately 17% for MedGemma 4B versus Gemma 3 4B (inaccurate ratings for 18.5% of Gemma 3 responses versus 15.4% of MedGemma responses; Supplementary Fig. 1). While the decrease in inaccurate responses for MedGemma suggests a benefit of the medically specialized model, this type of evaluation is subject to potential interrater variability and application-specific considerations, highlighting the importance of use-case-specific fine-tuning and evaluation.

MedGemma is also capable of engaging in open-ended clinical reasoning tasks, a relatively unique capability for a model of its size. We explored this capability using open-ended clinical vignette prompts and examined the MedGemma text output through clinical expert

Table 2 | Accuracy on medical text benchmarks

Model ^a	Open ^b	MedQA	MedMCQA	PubMedQA	MLLM anatomy	MLLM clinical knowledge	MLLM college biology	MLLM college medicine	MLLM medical genetics	MLLM professional medicine	MLLM virology	AfriMed-QA	MedXpertQA text only (OOD)	MedXpertQA multimodal (OOD)
Small models														
MedGemma 4B	✓	64.1	55.7	73.4	59.3	71.3	70.8	65.3	83.0	76.8	53.0	52.0	14.2	24.4
Gemma 3 4B	✓	52.4	45.7	68.4	54.1	70.2	77.8	63.6	74.0	65.4	42.8	48.0	11.6	22.3
MedGemma 27B Text ^c (with test-time scaling ³⁵)	✓	86.2	72.4	76.2	83.7	86.0	96.5	86.1	97.0	93.4	53.6	84.0	25.7	33.0^e
Gemma 3 27B	✓	74.9	62.6	73.4	74.8	86.0	93.8	78.6	91.0	85.7	51.2	68.0	15.7	29.8
BioMistral DARE 7B ^{6,d}	✓	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 (ref. 40) ^d	✓ ^e	61.4 ^f	61.2	74.2	71.9	78.1	82.6	71.1	83.0	78.7	N/A	N/A	-	-
OpenBioLLM 8B ^{7,d}	✓	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 ^{41,d}	✗	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 ^{7,d}	✓	78.2	74.0	79.0	83.9	92.9	93.8	85.7	93.2	93.8	N/A	N/A	-	-
DeepSeek-R1 (ref. 16)	✓	91.0	78.8	77.2	91.1	91.7	98.6	90.8	99.0	95.6	56.0	92.0	-	-
Gemini 2.5 Flash	✗	91.5	79.7	76.2	91.1	91.3	96.5	85.0	100.0	97.1	56.0	84.0	36.2	47.4
Gemini 2.5 Pro	✗	93.9	83.6	75.8	91.1	91.7	98.6	89.0	96.0	96.3	56.0	84.0	43.1	58.9
GPT 4o	✗	91.0	76.0	78.4	87.4	89.4	94.4	87.9	98.0	94.1	58.4	80.0	-	-
GPT o3	✗	93.3	83.9	80.8	92.6	94.7	99.3	90.2	100.0	96.7	56.0	84.0	54.6	67.5

Metrics for MedQA correspond to the original four-option test set, unless otherwise specified. Metrics for external small models as well as OpenBioLLM 70B were obtained from prior reports for these models. Other metrics were computed internally as described in the evaluation section. The MedXpertQA multimodal subset is the only benchmark that is not exclusively text and is included for completeness. Bold values indicate the best results in each section. A dash (-) indicates 'not calculated'. ^aSee Supplementary Note—General evaluation approach for inclusion criteria. ^bOpen-weight models. ^cMedGemma 27B Text performance is reported for all benchmarks except the MedXpertQA multimodal subset, on which MedGemma 27B performance is reported. ^dPrior reported results. ^eResearch use only. ^fResults for the all-options version of the dataset.

Table 3 | Accuracy on zero-shot medical image classification

Modality	Dataset	Metric	Small models				Large models		SoTA VLM ^a
			MedGemma 4B	Gemma 3 4B	MedGemma 27B	Gemma 3 27B	Gemini 2.5 Flash	Gemini 2.5 Pro	
CXR	MIMIC-CXR Med-Gemini test set ^b	Macro F1 (5)	88.9	81.2	90.0	71.7	81.0	85.8	90.7 Med-Gemini ⁹
	MIMIC-CXR MAIRA test set	Macro F1 (5)	40.5	26.7	40.2	25.0	32.2	31.9	51.6 Med-PaLM ^{M2}
	CheXpert (OOD)	Macro F1 (5)	48.1	32.6	49.9	26.2	37.4	37.0	49.0 ^c RadVLM ⁴³
	CXR14 (OOD) ^b	Macro F1 (3)	50.1	32.0	45.3	31.4	36.6	39.2	46.7 Med-Gemini ⁹
Histopathology	PathMCQA	Accuracy	69.8	37.1	71.6	42.2	46.9	42.7	–
Dermatology	US-Derm MCQA	Accuracy	71.8	52.5	71.7	66.9	78.4	81.0	–
Retina	EyePACS	Accuracy	64.9	14.4	75.3	20.3	17.5	27.7	–

Performance on CXR classification, formulated as a zero-shot generative task, and on histopathology, dermatology and retina image classification, formulated as zero-shot, multiple-choice tasks on internal datasets. For the MIMIC-CXR and CheXpert datasets, macro F1 is calculated for five conditions—atelectasis, cardiomegaly, consolidation, edema and pleural effusion. For CXR14, macro F1 is calculated for three conditions—lung opacity, pneumothorax and fracture. Bold values indicate the best results within the Small models section. Previously reported state-of-the-art (SoTA) results are also provided for comparison when available. ^aState-of-the-art at the time of MedGemma model release in July 2025, based on prior reported results. ^bRadiologist-adjudicated labels are used as described in ref. 9 and Supplementary Note—Evaluation parameters, so metrics may not be directly comparable to those reported in the literature for external models. ^cRadVLM macro F1 was calculated based on the individual F1 scores reported and estimated from the associated bar chart.

review. Overall, MedGemma effectively summarized key components of the scenarios and identified relevant management recommendations. Examples of these tasks, along with clinical commentary, are shown in Supplementary Note—Medical reasoning examples.

Medical visual question answering. To evaluate medical VQA, we used the often-reported benchmarks SLAKE²⁰ and VQA-RAD²¹. While different implementations for evaluating these benchmarks can make direct comparisons across literature challenging, we report direct comparisons of MedGemma to a number of additional models where possible as one component of demonstrating its strong baseline capabilities across modalities. Note that to avoid train–test contamination in VQA-RAD, we used the train and test splits described in refs. 9,22. These results are summarized in Extended Data Table 2. Again, MedGemma outperformed the standard Gemma 3 model and overall performed similarly to the much larger API-based models. LLaVA-Med and BiomedGPT-B are also included in the results table using the available reported metrics^{23,24}, noting that these models were involved in supervised fine-tuning (SFT) for the specific tasks and datasets. In addition to these quantitative results, we show several illustrative examples of open-ended visual question answering in Fig. 2.

It is important to clarify that we did not compare MedGemma with specialized models in the literature for which these metrics were calculated using a discriminative, embedding-based approach. Such approaches can be informative but use the training set to define the set of possible text answers, and thus represent the task as a limited-vocabulary classification task rather than an open-ended VQA task. Similarly, while LLaVA-Tri²⁵ also reports strong performance, the reported methodological details for these metrics are insufficient to support a direct comparison. Lastly, several other models have reported strong performance on these VQA benchmarks, such as HuatuoGPT-Vision²⁶ and PMC-Cambrian-8B-AN²⁷, with accuracy results reported in aggregate across open and closed questions. Given the challenge of comparing metrics even for the same benchmarks, we interpret the reported results for MedGemma on these VQA tasks as an important component towards demonstrating strong baseline performance across modalities.

CXR report generation. Results of performance for CXR report generation are shown in Table 4. These results are based on the MIMIC-CXR dataset and the RadGraph F1 metric²⁸. The pretrained checkpoint of MedGemma 4B demonstrates results similar to or better than other top-performing models, which is particularly notable given the comparator models are either larger or specialized only for this task. This shows the versatility of the Gemma architecture in adapting to complex

medical use cases and enabling MedGemma. In addition, expert human evaluation demonstrated that 68% of generated reports from normal studies and 49% of generated reports from abnormal studies were rated as equal to or superior to the original reports (Extended Data Fig. 1). This corresponds to 81% of generated reports rated as likely to result in the same or superior clinical decisions as compared to the original reports. This represents superior performance to the larger Med-Gemini model, for which 73% of generated reports were rated as resulting in the same or superior clinical decisions as compared to the original reports⁹.

Again, the open nature of MedGemma has enabled third-party evaluations. Specifically, for CXR report generation, the ReXrank leaderboard (ReXrank.ai) provides standardized evaluation across several models, datasets and metrics for CXR interpretation. Notably, MedGemma 4B demonstrates the highest Generative Radiology Report Evaluation and Error Notation (GREEN) score²⁹ on the ReXgradient dataset³⁰ (GREEN score 0.566) and the IU-Xray set³¹ (GREEN score 0.724), with middle-of-the-leaderboard performance on the MIMIC-III and CheXpert Plus datasets. MedGemma also demonstrates top-three performance on the ReXVQA medical VQA evaluation³². (ReXrank.ai, accessed 9 February 2026). While these results are consistent with strong baseline performance for MedGemma on CXR report generation and VQA, including on datasets not seen during training, the thorough ReXrank evaluation also notably illustrates that different models can excel on different metrics and datasets. As such, users can consider the aspects and metrics most important for their use case when selecting a model for further development and validation.

CXR anatomical structure localization. Although specific optimizations were not performed for this task, because anatomical bounding boxes were included in the training data for MedGemma 27B, we evaluated the localization of anatomical structures using the intersection over union (IoU) metric on the test set of the Chest ImaGenome dataset³³. The findings are included in Supplementary Table 9 for reference, with IoU values of 0.16 for MedGemma 27B, 0.08 for baseline Gemma 3 27B, 0.03 for MedGemma 4B and 0.09 for Qwen3-VL 4B. These results demonstrate improved performance on this task when relevant training data are added, but indicate there is still substantial room for optimization and improvement of this capability.

Medical agentic behavior. To provide initial insights into the capabilities of MedGemma in the setting of use cases beyond single-turn question answering, we also evaluated MedGemma in an agentic role using the AgentClinic benchmark¹⁵. In this task, the model functions as a ‘physician agent’ in simulated clinical encounters, performing tasks such as patient history-taking, exam interpretation and diagnosis

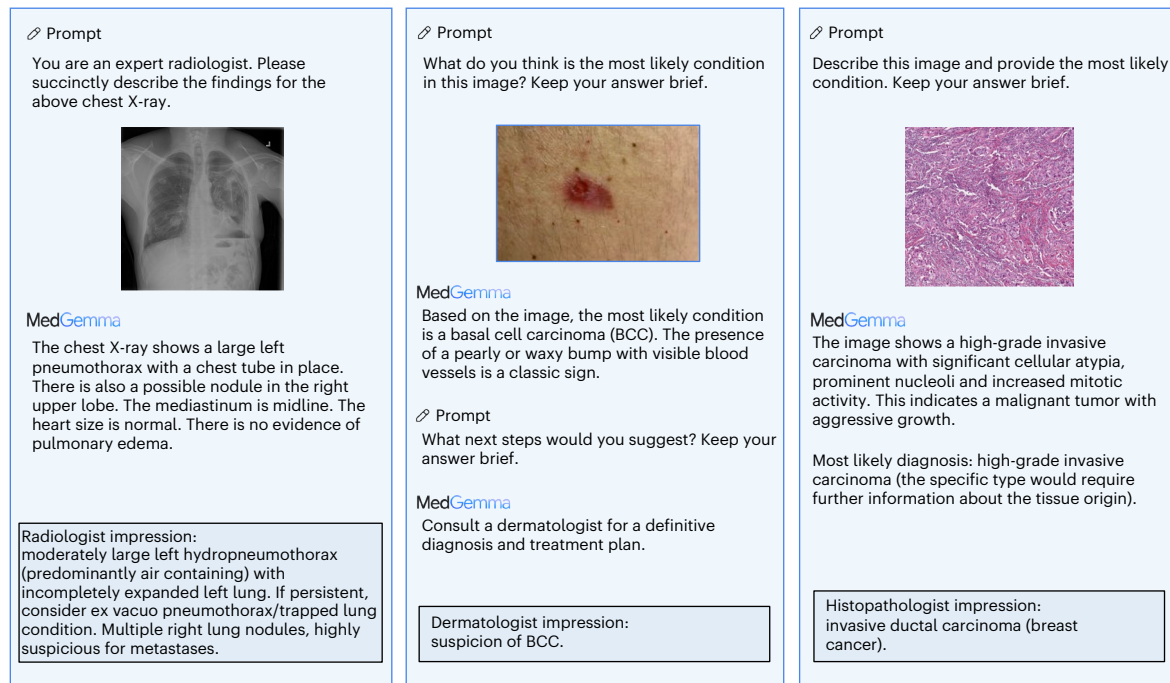


Fig. 2 | Examples of open-ended multimodal question answering with MedGemma 4B. Prompts that include both text and an image are shown along with the MedGemma 4B response. Assessments of the response by US board-certified clinicians are shown at the bottom of each section. X-ray image sourced from the NIH Clinical Center (<https://nihcc.app.box.com/v/ChestXray-NIHCC/>

[folder/37178474737](https://nihcc.app.box.com/v/ChestXray-NIHCC/)), reproduced from ref. 54; dermatology image sourced from the Skin Condition Image Network dataset (<https://github.com/google-research-datasets/scin>), reproduced from ref. 55; and histopathology image sourced from The Cancer Genome Atlas⁵⁶.

under uncertainty to reach the correct answers. Using the text-only AgentClinic-MedQA and AgentClinic-MIMIC-IV environments (with GPT 4o for the patient agent and measurement), MedGemma 27B exceeded the performance of human physicians on AgentClinic-MedQA and approached the performance of much larger models on both datasets. Results are summarized in Extended Data Table 3. Notably, we found that the smaller Gemma 3 4B and MedGemma 4B models were poorly suited to this task, demonstrating difficulty following system instructions within the AgentClinic agentic framework.

General purpose benchmarks. Given the limitations and poor performance that many medically specialized models exhibit when faced with nonmedical tasks, which may critically interweave with medical applications, we evaluated possible trade-offs of specialization. Models were evaluated on general language model benchmarks including MMLU Pro, Global MMLU Lite and MMMU datasets, with performance comparisons between MedGemma and Gemma 3 variants shown in Extended Data Table 4. MedGemma demonstrated only minor decreases in performance relative to the general models of the same size, further suggesting its potential utility for downstream applications that require both specialized and 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 often performed better, medically specialized models such as MedGemma exhibited a distinct performance-efficiency advantage when their substantially smaller size was taken into account. This size-to-performance benefit was especially pronounced in medical vision-based tasks, where MedGemma surpassed the performance of substantially 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 for model development and evaluation.

Fine-tuning MedGemma enables robust task-specific adaptation

While the MedGemma models achieved strong baseline performance, these models are intended as a useful starting point for specialized task adaptation (for example, specific styles, specific classification tasks, or other user needs).

We conducted fine-tuning experiments using three different datasets (not previously included in MedGemma training) and using either MedGemma 4B or Gemma 3 4B as the base model. The results are shown in Extended Data Fig. 2 and generally demonstrate an increase in performance with fine-tuning and with higher performance for MedGemma 4B as compared to Gemma 3 4B. This performance advantage for MedGemma was particularly notable when using only 10% of the available data for a given task, supporting the notion of MedGemma as a useful starting point for data-efficient fine-tuning. We also note that for SIIM-ACR, Gemma 3 collapses to predicting the majority class, whereas MedGemma avoids this collapse, likely due, at least in part, to being pretrained on radiology data. While these results are provided to demonstrate illustrative comparisons of fine-tuning for classification tasks using a common SFT approach, specific fine-tuning methods and prompt engineering may benefit from further use-case-specific optimization.

Fine-tuning of MedGemma 4B with reinforcement learning (RL) was also performed to demonstrate further optimized performance for CXR report generation, as a task that was also included in MedGemma training. The resulting fine-tuned model achieved a RadGraph F1 score of 30.3 (for findings plus impression) on MIMIC-CXR, representing state-of-the-art performance for this specific metric and dataset, as shown in Table 4.

Table 4 | Report generation on the MIMIC-CXR dataset

Model	Sections	Clinical metric (%) RadGraph F1
CXR-RePair ⁴⁴	F only	9.1
M ² transformer ⁴⁵	F only	22.0
Med-PaLM M, 12B ⁴²	F only	25.2
Med-PaLM M, 84B ⁴²	F only	26.7
MAIRA-1 (ref. 46)	F only	24.3
MAIRA-2 (ref. 47)	F only	34.6
R2Gen ⁴⁸	F+I	13.4
WCT ⁴⁹	F+I	14.3
CvT-21DistillGPT2 (ref. 50)	F+I	15.4
Flamingo-CXR ⁵¹	F+I	20.5
Med-Gemini-2D ⁹	F+I	24.4
PaliGemma 2 10B ⁵²	F+I	29.5
MedVersa ⁵³	F+I	30.0
MedGemma 4B PT	F+I	29.5
MedGemma 4B with RL	F+I	30.3

The ‘sections’ column indicates whether the model generates the findings (F) only or both the findings and impression (F+I) sections of the report; generation of findings and impressions is generally considered a more difficult problem as the impression section includes the radiologist’s comprehensive interpretation of the study. Previously reported RadGraph F1 metrics were sourced from existing published research except for MedGemma 4B PT and MedGemma 4B with task-specific RL, which were computed in this work. Higher scores are better. Bold values indicate the best results in each section. Please note that scores across models may not be directly comparable due to differences in inclusion and exclusion criteria across studies. PT, pretrained.

We also performed RL and evaluation experiments on EHRQA, a synthetic dataset specifically focused on longitudinal outpatient electronic health record data. Fine-tuning MedGemma 27B Text using RL on the EHRQA training set resulted in improved performance, with the greatest gains observed for question categories requiring reasoning across interdependent records. These results further support the potential of fine-tuning MedGemma for effective use in task-specific electronic health record applications. Additional details about the dataset and experiments are described in the Supplementary Note.

Discussion

We introduce MedGemma, a new collection of medical vision-language foundation models, and MedSigLIP, a multidomain medical image encoder. These models were built on Gemma 3, with optimization for medical domains. We evaluated across a range of medical benchmarks that include tasks in clinical reasoning, biomedical knowledge, report generation and medical image classification, finding strong performance for MedGemma and MedSigLIP. Performance could be further improved through fine-tuning, highlighting the flexible nature of these models and their potential as useful starting points for developing AI applications in healthcare. The development of a vision encoder that can effectively tokenize across different medical image modalities combined with effective alignment for a performant language model represents key contributions of this work. As one concrete example of the utility of a single model that is capable across multiple modalities, developers have tuned MedGemma across different modalities using a unified pipeline, allowing more efficient and effective integration into a single system for general clinical practice guideline retrieval and navigation³⁴.

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 the optimized incorporation of domain-specific data for both

pretraining and post-training, as illustrated by improvements over base Gemma 3 models across all evaluated benchmarks and the achievement of performance on par with much larger models.

Compared with large or API-based frontier models, MedGemma offers advantages when the use case requires any of the following: a frozen model for documentation and reliability, sensitivity to training or inference costs, the ability to run locally or offline, specific medical image and text capabilities, or full control over model adaptation. The frontier models, on the other hand, may still be the preferred choice when the user requires optimal broad performance without the above constraints. Notably, large models may also be used in concert with models such as MedGemma in agentic settings to achieve useful combinations of capabilities.

The MedGemma collection of models enables a wide range of potential downstream applications for the developer community, with some examples described below. The multimodal capabilities, including access to image and text embeddings, may be particularly useful for medical image retrieval. This could aid interpretation by referencing similar past cases, as well as enable the development of research cohorts and the creation of educational tools. MedGemma enables 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 and 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 serve as powerful tools within agentic frameworks, combining capabilities 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 to provide a snapshot of their capabilities. However, we note that limitations exist for these benchmarks such that evaluating performance on them represents only the first step towards validating real-world utility^{35,36}. Additionally, some benchmarks may be near saturation in terms of model performance, 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, for example, in ref. 37. More work is also needed to understand the performance capabilities and requirements for 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 also note that, across modalities, specialized models may still provide better performance or be better suited to some tasks. For example, while MedSigLIP and MedGemma demonstrate performance on par with many histopathology-specialized models for some tasks, state-of-the-art vision encoders for histopathology demonstrate superior performance on several benchmark tasks. Further optimization of pathology performance represents an area for possible improvement. Additional important areas for possible development include extending radiology capabilities to include three-dimensional series for CT and MRI, as well as capabilities for additional modalities such as genomics. Practitioners should compare MedGemma to other base models on their task of interest when possible. At the same time, while a specialized model might score higher on some tasks, a key strength of a multimodal medical foundation model such as MedGemma lies in its flexibility. Although MedGemma surpasses frontier model performance on some tasks (Table 3), its value also includes aspects other than performance, such as enabling data and model sovereignty, as well as computational efficiency. Instead of requiring separate adaptation and implementation approaches for each specific disease or image

type, a unified approach can be applied to many tasks. This flexibility also extends to potential utility for use cases that benefit from language generation and multiple image modalities, enabling systems that can adapt to new medical challenges and help connect different aspects of patient information.

We openly released MedGemma and MedSigLIP to facilitate their widespread evaluation, improvement and adaptation by the community across multiple modalities. Initial examples from the broader research community of research performed on and with MedGemma following its initial release are summarized in Extended Data Table 5, including evaluations of OOD modalities and tasks. In addition to enabling evaluation and reproducibility, MedGemma's openness provides developers with predictability and flexibility for extensive model adaptation and validation in real-world healthcare applications. We hope that this approach accelerates the development of AI applications across a broad array of healthcare use cases.

Conclusion

In this study, we showed that MedGemma models demonstrate strong baseline performance across a variety of vision-language and text-only medical tasks, as well as the ability to further improve performance through task-specific fine-tuning. We also showed that MedSigLIP demonstrates robust multidomain capabilities and can thus serve as a strong medical vision encoder foundation model. These results demonstrate that optimizing approaches across pretraining and post-training can enable multimodal capabilities across several medical imaging domains, providing a unified starting point for further development. The breadth and efficiency of these models offer exciting possibilities to address a range of use cases. At the same time, the potential for model errors or subtle forms of distribution bias in training data remains an important aspect to address through continued model improvements, mitigation strategies and ongoing validation³⁸. Thoughtful validation of safety, performance and reliability for downstream applications remains a critical aspect to advancing the use of multimodal AI models in medicine. By providing MedGemma and MedSigLIP models to the developer community with transparent methodologies, we hope to see them enable many useful and innovative medical applications.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-026-04626-w>.

References

- Alayrac.J.-B. 2022. Flamingo: a visual language model for few-shot learning. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 23716–23736 (Curran Associates Inc., 2022).
- Achiam, J. et al. GPT-4 technical report. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2303.08774> (2023).
- Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2302.13971> (2023).
- Toma, A. et al. Clinical Camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2305.12031> (2023).
- Han, T. et al. MedAlpaca—an open-source collection of medical conversational AI models and training data. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2304.08247> (2023).
- Labrak, Y. et al. BioMistral: a collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 5848–5864 (2024).
- Gemma-Team et al. Gemma 3 technical report. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2503.19786> (2025).
- Saab, K. et al. Capabilities of Gemini models in medicine. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2404.18416> (2024).
- Yang, L. et al. Advancing multimodal medical capabilities of Gemini. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2405.03162> (2024).
- Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
- Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning* (eds Flores, G. et al.) 248–260 (PMLR, 2022).
- Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: a dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (eds Inui, K. et al.) 2567–2577 (2019).
- Hendrycks, D. et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (2021).
- Olatunji, T. et al. AfriMed-QA: a pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (eds Che, W. et al.) 1948–1973 (2024).
- Schmidgall, S. et al. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2405.07960> (2024).
- DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *Nature* **645**, 633–638 (2025).
- Ankit Pal, M. S. OpenBioLLMs: advancing open-source large language models for healthcare and life sciences <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B> (2024).
- Warner, B. et al. Medmarks: a comprehensive open-source LLM benchmark suite for medical tasks. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2605.01417> (2026).
- Pandit, S. et al. MedHallu: a comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (eds Christodoulopoulos, C. et al.) 2858–2873 (2025).
- Liu, B. et al. SLAKE: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1650–1654 (IEEE, 2021).
- Lau, J. J., Gayen, S., Ben Abacha, A. & Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **5**, 180251 (2018).
- Xu, S. et al. ELIXR: towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2308.01317> (2023).
- Li, C. et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36**, 28541–28564 (2023).
- Zhang, K. et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. Preprint at *arXiv* <https://arxiv.org/abs/2305.17100v2> (2024).
- Xie, Y. et al. MedTrinity-25M: a large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations* (2025).

26. Chen, J. et al. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y. -N.) 7346–7370 (2024).
27. Wang, J. et al. SemiHVision: enhancing medical multimodal models with a semi-human annotated dataset and fine-tuned instruction generation. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2410.14948> (2024).
28. Jain, S. et al. RadGraph: extracting clinical entities and relations from radiology reports. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2106.14463> (2021).
29. Ostmeier, S. et al. GREEN: Generative radiology report evaluation and error notation. *Findings of the Association for Computational Linguistics: EMNLP 2024* 374–390 <https://doi.org/10.18653/v1/2024.findings-emnlp.21> (Association for Computational Linguistics, 2024).
30. Zhang, X., Acosta, J. N., Miller, J., Huang, O. & Rajpurkar, P. ReXGradient-160K: a large-scale publicly available dataset of chest radiographs with free-text reports. In *Machine Learning for Health* (2025).
31. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).
32. Pal, A. & Sankarasubbu, M. Gemini goes to Med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proc. 6th Clinical Natural Language Processing Workshop* (eds Naumann, T. et al.) 21–46 (2024).
33. Wu, J. et al. Chest ImaGenome dataset. *PhysioNet*. <https://doi.org/10.13026/wv01-y230> (2021).
34. Zun, L. Q., Halim, M. Z. B. A. & Fye, G. M. Fine-tuning MedGemma for clinical captioning to enhance multimodal RAG over Malaysia CPGs. In *IEEE 13th Conference on Systems, Process & Control (ICSPC)* 129–134 (2025).
35. Alaa, A. et al. Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track* (2025).
36. Mahmood, F. A benchmarking crisis in biomedical machine learning. *Nat. Med.* **31**, 1060 (2025).
37. Bedi, S. et al. MedHELM: holistic evaluation of large language models for medical tasks. *Nat. Med.* **32**, 943–951 (2026).
38. Kim, Y. et al. Medical hallucinations in foundation models and their impact on healthcare. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2503.05777> (2025).
39. Muennighoff, N. et al. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (eds Christodoulopoulos, C. et al.) 20275–20321 (Association for Computational Linguistics, 2025).
40. John Snow Labs. JSL-MedLLaMA-3-8B-v2.0 <https://huggingface.co/johnsnowlabs/JS-MedLlama-3-8B-v2.0> (accessed 11 June 2025).
41. IQVIA. Introducing IQVIA Medical Reasoning (Med-R1 8B): Best-in-class medical reasoning LLM. <https://www.iqvia.com/blogs/2025/04/introducing-iqvia-medical-reasoning-med-r1-8b> (accessed 11 June 2025).
42. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* **1**, A10a2300138 (2024).
43. Deperrois, N. et al. RadVLM: a multitask conversational vision-language model for radiology. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2502.03333> (2025).
44. Endo, M., Krishnan, R., Krishna, V., Ng, A. Y. & Rajpurkar, P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of the Machine Learning for Health* (eds Roy, S. et al.) Vol. 158, 209–219 (PMLR, 2021).
45. Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P. & Jurafsky, D. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Toutanova, K. et al.) 5288–5304 (2021).
46. Hyland, S. L. et al. MAIRA-1: a specialised large multimodal model for radiology report generation. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2311.13668> (2023).
47. Bannur, S. et al. MAIRA-2: grounded radiology report generation. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2406.04449> (2024).
48. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (eds Webber, B. et al.) 1439–1449 (2020).
49. Moens, M.-F., Huang, X., Specia, L., Yih, & S. W.-T. (eds). Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* 4009–4015 (Association for Computational Linguistics, 2021).
50. Nicolson, A., Dowling, J. & Koopman, B. Improving chest x-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **144**, 102633 (2023).
51. Tanno, R. et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat. Med.* **31**, 599–608 (2025).
52. Steiner, A. et al. PaliGemma 2: a family of versatile VLMs for transfer. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2412.03555> (2024).
53. Zhou, H.-Y., Adithan, S., Acosta, J. N., Topol, E. J. & Rajpurkar, P. MedVersa: a generalist learner for multifaceted medical image interpretation. *NEJM AI* **3**, A10a2500595 (2026).
54. Wang, X. et al. 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* 2097–2106 (IEEE, 2017).
55. Ward, A. et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Netw. Open* **7**, e2446615 (2024).
56. Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

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

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

© The Author(s) 2026

Andrew Sellergren^{1,4}✉, Sahar Kazemzadeh^{1,4}, Fereshteh Mahvar¹, Atilla Kiraly¹, Madeleine Traverse¹, Timo Kohlberger¹, Chufan Gao¹, Ronnachai Jaroensri¹, Shawn Xu¹, Fayaz Jamil¹, Cian Hughes¹, Charles Lau^{1,2}, Justin Chen¹, Liron Yatziv¹, Tiffany Chen¹, Bram Sterling¹, Stefanie Anna Baby¹, Susanna Maria Baby¹, Jeremy Lai¹, Samuel Schmidgall¹, Lu Yang¹, Kejia Chen¹, Per Bjornsson¹, Shashir Reddy¹, Ryan Brush¹, Kenneth Philbrick¹, Bashir Sadjad¹, Mercy Asiedu¹, Ines Mezerreg¹, Howard Hu¹, Howard Yang¹, Richa Tiwari¹, Sunny Jansen¹, Preeti Singh¹, Yubin Kim¹, Daniel McDuff¹, Catherine Kozlowski³, Rachelle Sico¹, Aishwarya Kamath¹, Johan Ferret¹, Shreya Pathak¹, Nino Vieillard¹, Ramona Merhej¹, Sarah Perrin¹, Tatiana Matejovicova¹, Alexandre Ramé¹, Morgane Riviere¹, Louis Rouillard¹, Thomas Mesnard¹, Geoffrey Cideron¹, Jean-bastien Grill¹, Sabela Ramos¹, Edouard Yvinec¹, Michelle Casbon¹, Elena Buchatskaya¹, Jean-Baptiste Alayrac¹, Dmitry Lepikhin¹, Vlad Feinberg¹, Sebastian Borgeaud¹, Alek Andreev¹, Cassidy Hardin¹, Robert Dadashi¹, Léonard Hussenot¹, Armand Joulin¹, Olivier Bachem¹, Clement Farabet¹, Joelle Barral¹, Tris Warkentin¹, Jonathon Shlens¹, David Fleet¹, Shekoofeh Azizi¹, Victor Cotruta¹, Omar Sanseviero¹, Gus Martins¹, Phoebe Kirk¹, Anand Rao¹, Yossi Matias¹, Katherine Chou¹, Avinatan Hassidim¹, Dale Webster¹, Shravya Shetty^{1,5}, Kavi Goel¹, Yun Liu¹, Can Kirmizibayrak¹, David F. Steiner^{1,5}✉, Rory Pilgrim^{1,5}, Daniel Golden^{1,5}✉ & Lin Yang^{1,5}✉

¹Google, Mountain View, CA, USA. ²Google via Advanced Clinical, Deerfield, IL, USA. ³Google via Buchanan Ingersoll and Rooney PC, Pittsburgh, PA, USA.

⁴These authors contributed equally: Andrew Sellergren, Sahar Kazemzadeh. ⁵These authors jointly supervised this work: Shravya Shetty, David F. Steiner, Rory Pilgrim, Daniel Golden, Lin Yang. ✉e-mail: asellerg@google.com; davesteiner@google.com; dangolden@google.com; linyan@google.com

Methods

Datasets

For general-purpose data replay during pretraining, the original data mixtures from SigLIP and Gemma 3 (ref. 7) were leveraged. The medical training and evaluation datasets largely followed those used in Med-Gemini⁹. Here we outline specific changes and differences in the datasets relative to the details described in ref. 9.

Training datasets. Text-only datasets. For text datasets, we sampled responses and logits from a large IT (instruction-tuned) teacher using the training splits of multiple medical question-answering datasets, including MedQA¹⁰, MedMCQA¹¹, PubMedQA¹², MedExpQA⁵⁷, AfriMed-QA¹⁴, HealthSearchQA⁵⁸, and LiveQA⁵⁹. We also sampled responses and logits for approximately 200,000 synthetic medical questions. To generate these synthetic questions, we prompted Gemini with five randomly selected questions from the MedQA training split and asked for new questions based on, but nonoverlapping with, the provided examples. After generating an initial set of 500,000 questions, we reduced the final set to 200,000 with the goal of maintaining question diversity. This process involved vectorizing the questions (using TF-IDF with document frequency between 0.5 and 5), reducing dimensionality to 100 features, generating 100 clusters using *k*-means, and finally sampling an equally distributed set from these clusters.

Multimodal datasets. The multimodal capabilities of MedGemma are focused on two-dimensional medical images (for example, X-rays, two-dimensional slices from CT and MRI). Relative to Med-Gemini⁹, training on two-dimensional medical images used the same data sources, while the three-dimensional volumes and genomic datasets described in ref. 9 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 because it focuses on six-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 included only 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 retina images), dermatology (51,049 more dermatology images with 210 different skin conditions), histopathology (a total of ~32.5 million patch–text pairs, downsampled during training) and radiology data (54,573 more CT 2D slices, 47,622 MRI 2D slices); see Extended Data Table 1. Additional details about the data sources comprising histopathology and radiology (CT-US1 and MRI-US1), along with the sampling and curation strategy, are described previously in ref. 9. CT and MRI slices used for training were curated based on text from the radiology reports specifying the location (that is, slice) of specific findings and representing the diversity of studies from a US-based radiology outpatient diagnostic center network. Histopathology patches were sampled across several datasets and tissue types using slide-level metadata and patch-level annotation to generate captions. Additional details on the datasets and captions are provided in Supplementary Tables 7 and 8. The internal dermatology dataset consists of de-identified dermatology images (both clinical and dermatoscopic) from Columbia, Australia and the US across tel-dermatology conditions, skin cancer, and normal skin, respectively. The Chest ImaGenome³³ and EHRQA (Supplementary Note—Additional evaluation using an internal synthetic EHR dataset: EHRQA) datasets were used only for post-training of MedGemma 27B as additional areas of exploration for this model variant.

Data preprocessing. Our data preparation followed the study in ref. 9 closely. Image padding and resizing algorithms remained 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 used

the SentencePiece tokenizer with 262,000 entries. Additionally, for CT images, we preselected the following three windows and converted them into the red, green and blue (RGB) color channels of the input image to highlight (1) bone and lung, window width of 2,250, window level of -100; (2) soft tissue, window width of 350, window level of 40; (3) brain, window width of 80, window level of 40.

Modeling methodology

Modeling architecture and training infrastructure. The MedGemma model architecture follows Gemma 3 (ref. 7) and is compatible with all existing Gemma infrastructure. The Gemma 3 vision encoder is the 400M variant of the SigLIP encoder and is shared across the different Gemma language model sizes (4B and 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 precomputed visual tokens for memory saving and used data and model sharding for multipod training.

Model training. The MedGemma multimodal models (MedGemma 4B and MedGemma 27B) used all of the following steps, while MedGemma 27B text leveraged the post-training stage alone (as we observed stronger text-only performance in initial tuning experiments with this approach).

Vision encoder enhancement for MedGemma. To improve the vision encoder's ability to encode and distinguish subtle differences in medical images, we fine-tuned the vision encoder in Gemma 3 (SigLIP-400M) using over millions of medical image–text pairs as listed in Extended Data Table 1. To retain SigLIP's existing performance, its original training data (for example, WebLI) were retained and medical data was mixed with a 2% weight during training.

To balance general-domain capabilities with specialized medical knowledge, the total proportion of medical data in the training mixture was constrained to 2%. Within this medical subset, data availability across modalities varied widely in scale. To fully use available data without inducing modality-specific overfitting, we used an iterative, empirical sampling strategy to define the final mixture weights. Specifically, sampling weights were optimized by tracking validation performance across a limited number of modality-specific tasks—CXR condition classification, pathology patch classification, diabetic retinopathy classification, dermatology condition classification and ImageNet zero-shot classification performance for general capability. Initial weights were set proportionally to the total number of examples per dataset, while limiting the total amount of medical data for training. After an initial training run, we analyzed the validation curves to adjust the distribution—for datasets on which the model overfit, weights were downsampled to match the number of examples seen at the step of peak performance. Conversely, for datasets on which validation metrics were still improving at the end of training, the sampling weights were doubled. This process was iterated until validation metrics across all modalities reached peak performance simultaneously toward the end of the training phase. The final optimized sampling distribution is detailed in Supplementary Table 5.

While the Gemma 3 vision encoder works at 896 × 896 resolution, we found that many medical vision tasks worked reasonably well at 448 × 448 resolution (Supplementary Table 2). Thus, while the MedGemma 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 downsampled positional embeddings to work with fewer input patches at the lower resolution.

Multimodal decoder pretraining. After the vision encoder enhancement, the Gemma language model needed to be readapted to the new vision encoder, not only for medical data but also for the general image domain to preserve its visual-language reasoning capabilities. This goal was achieved in the pretraining stage by 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 (Extended Data Table 1) with 12% weight (based on tuning experiments), trained for approximately five epochs on the medical mixture given the mixing ratio, and selected the checkpoint based on the validation set performance on CXR report generation and radiology, dermatology and ophthalmology visual question answering.

For multimodal decoder training, the total proportion of medical data in the training mixture was constrained to approximately 12%. An empirical sampling strategy similar to that used for vision encoder enhancement was used to define the final mixture weights, with tuning performance across a limited number of modality-specific tasks: RadGraph F1 on MIMIC-CXR (radiology), ROUGE-L for patch captioning (pathology) and Top-1 classification accuracy for both dermatology (skin conditions) and ophthalmology (retinal imaging). As above, weights were adjusted based on peak performance or ongoing improvement at the end of the training run. The final optimized sampling distribution is detailed in Supplementary Table 6, ensuring that high-volume datasets were used effectively without dominating the training distribution or causing early overfitting.

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 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) RL—medical imaging data with paired text was used in the RL stage of post-training. For multimodal training, we found that RL enabled better generalization than SFT, so all multimodal post-training was performed via RL.

MedGemma fine-tuning. To demonstrate MedGemma’s capacity for improved task-specific performance relative to a nonmedical base model, we conducted fine-tuning experiments comparing use of MedGemma 4B and Gemma 3 4B as base models. These experiments included the following three multimodal tasks involving previously unseen datasets: (1) pneumothorax classification on CXRs via SIIM-ACR⁶⁰, (2) histopathology patch classification via CRC100k⁶¹ and (3) skin lesion classification via ISIC 2017 (ref. 62). The base models were fine-tuned through SFT using low-rank adaptation using either all training data available (100%) or a 10% class-balanced selection to evaluate data efficiency. For SIIM-ACR and CRC100k, because these datasets provide no validation (tuning) splits, we reserved a random 10% per label of the original training data to form a validation set. For each task, we used a constant learning rate of 4×10^{-5} and fine-tuned the model up to four epochs on the full training dataset. Checkpoints that achieved the highest performance on the validation set were selected for the final evaluation on the test set. The balanced accuracy metric was selected to account for imbalanced label distribution in the test sets. Confidence intervals were generated by bootstrap resampling with 1,000 samples. Additionally, we performed RL-based fine-tuning for MIMIC-CXR report generation, starting with MedGemma 4B-IT as the base model, using the image and indication sections of the report as input, and training the model to complete the findings and impression sections.

Model evaluation. Model evaluation details for the different tasks are provided in Supplementary Note—MedGemma evaluations. Unless otherwise reported, all evaluations consisted of a single inference run per example. For MedGemma evaluations, a temperature of 0.0 was used for medical question sets and the default temperature was used on non-medical benchmarks. For evaluations of all other models, on all datasets, each model’s default temperature and top-*k* settings were used. Additionally, details of the evaluation of the MedSigLIP encoder can be found in Supplementary Note—MedSigLIP evaluations. Evaluation using datasets completely external to training of the vision encoder, pretraining, or post-training is specified as OOD evaluation. All evaluations were conducted using data sources or data splits that were completely held out from MedGemma-specific training. Many members of the research community have also evaluated MedGemma and MedSigLIP in various scenarios following the release of the models^{34,63–70} and these results are summarized in Extended Data Table 5.

Use of AI in paper preparation

The introduction section of this paper was drafted manually and then suggestions from Gemini were reviewed and incorporated where useful. Final manual checks were performed to ensure content accuracy. The authors take full responsibility for the content.

Reporting summary

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

Data availability

De-identified data from public and private datasets were used. Detailed datasets are listed in Table 1 and Extended Data Table 1. Due to licensing restrictions, we could not share the private datasets, including EHRQA, US-Derm MCQA, PathMCQA, CT-US1, MRI-US1, internal histopathology, internal dermatology and EyePACS. Other training and evaluation datasets are publicly available.

Code availability

The models have been released openly, with information and details at the Google Health AI Developer Foundations site at <https://goo.gle/hai-def> and <https://goo.gle/medgemma>. These public sites also include code examples for running, fine-tuning and evaluating MedGemma.

References

- Alonso, I., Oronoz, M. & Agerri, R. MedExpQA: multilingual benchmarking of large language models for medical question answering. *Artif. Intell. Med.* **155**, 102938 (2024).
- Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
- Abacha, A. B., Agichtein, E., Pinter, Y. & Demner-Fushman, D. Overview of the medical question answering task at TREC 2017 LiveQA. <https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf> (2017).
- Zawacki, A. et al. SIIM-ACR pneumothorax segmentation. <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation> (2024).
- Kather, J. N., Halama, N. & Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue (version v0.1) [dataset]. Zenodo <https://doi.org/10.5281/zenodo.1214456> (2018).
- Codella, N. C. F. et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In *IEEE 15th International Symposium on Biomedical Imaging*, 168–172 (2018).
- Pal, A. et al. ReXVQA: a large-scale visual question answering benchmark for generalist chest X-ray understanding. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2506.04353> (2025).

64. Wu, J. et al. BRIDGE: benchmarking large language models for understanding real-world clinical practice text. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-026-01719-2> (2026).
 65. Baharoon, M. et al. RadGame: an AI-powered platform for radiology education. In *Proc. 5th Machine Learning for Health Symposium* (eds Argaw, P. et al.) Vol. 297, 898–920 (2026).
 66. Zhou, S. et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digit. Med.* **9**, 34 (2026).
 67. Sung, N.-J., Lee, D., Choi, B. H. & Park, C. J. AMRG: extend vision language models for automatic mammography report generation. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2508.09225> (2025).
 68. Gautam, S., Riegler, M. A. & Halvorsen, P. Kvasir-VQA-x1: a multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy. In *Data Engineering in Medical Imaging* (eds Bhattarai, B. et al.) 53–63 (2025).
 69. Aali, A. et al. MedVAL: toward expert-level medical text validation with language models. *npj Digit. Med.* <https://doi.org/10.1038/s41746-026-03084-5> (2026).
 70. Cui, X., Huang, T., Wang, Y., Hu, L. & Yin, L. MedFM-Robust: benchmarking robustness of medical foundation models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2605.19027> (2026).
- the data. A.S., S.K., F.M., A. Kiraly, M.T., T.K., C.G., R.J., S.X., L. Yatziv, T.C., B. Sterling, S.A.B., S.M.B., J.L., S. Schmidgall, Lu Yang, K. Chen, P.B., S. Reddy, R.B., K.P., B. Sadjad, M.A., I.M., Y.K., D.M., A. Kamath, J.F., S. Pathak, N.V., R.M., S. Perrin, T. Matejovicova, A. Ramé, M.R., L.R., T. Mesnard, G.C., J.-B.G., S. Ramos, E.Y., M.C., E.B., J.-B.A., D.L., V.F., S.B., A.A., C. Hardin, R.D., L.H., A.J., O.B., C.K. and Lin Yang contributed to the technical implementation. C. Hughes, C.L., J.C. and D.F.S. contributed to clinical guidance. C.F., J.B., T.W., J.S., D.F., Y.M., K. Chou, A.H., D.W., S. Shetty, K.G. and Y.L. provided strategic guidance. A.S., F.J., H.H., H.Y., R.T., S.J., P.S., D.F.S., R.P., D.G. and Lin Yang contributed to paper organization and team logistics. F.M., L. Yatziv, T.C., B. Sterling, S.A.B., S.M.B., K.P., V.C., O.S., G.M., P.K., A. Rao, C. Kirmizibayrak, R.S. and C. Kozlowski contributed to externalization of the open-source infrastructure. A.S., S.K., A. Kiraly, M.T., T.K., C.G., S.A., D.F.S., D.G. and Lin Yang contributed to drafting and revising the paper.

Acknowledgements

Many teams from both Google Research and Google DeepMind collaborated extensively on this project. The authors thank E. Wulczyn for their feedback and insights, which substantially enhanced this paper; N. Hammel, L. Foster and K. Parakh (Google colleagues) for their review of the qualitative examples shown in this paper; and A. Steiner and X. Wang (Google colleagues) for sharing their expertise with the Big Vision infrastructure. The results shown here are in part based on data generated by The Cancer Genome Atlas Research Network. The NIH CXR images are provided by the NIH Clinical Center and can be downloaded from <https://nihcc.app.box.com/v/ChestXray-NIHCC>. Dermatology data from the Skin Condition Image Network dataset are subject to the Skin Condition Image Network license, available at <https://github.com/google-research-datasets/scin/blob/main/LICENSE>.

Author contributions

A.S., S.K., D.F.S., R.P., D.G. and Lin Yang conceptualized and designed the study. F.J., D.F.S., R.P. and D.G. contributed to the acquisition of

Funding statement

This study was funded by Alphabet and/or a subsidiary thereof ('Alphabet').

Competing interests

All authors, except for C.L. and C. Kozlowski, are current or former employees of Google LLC, a subsidiary of Alphabet, and may own stock as part of their standard compensation packages.

Additional information

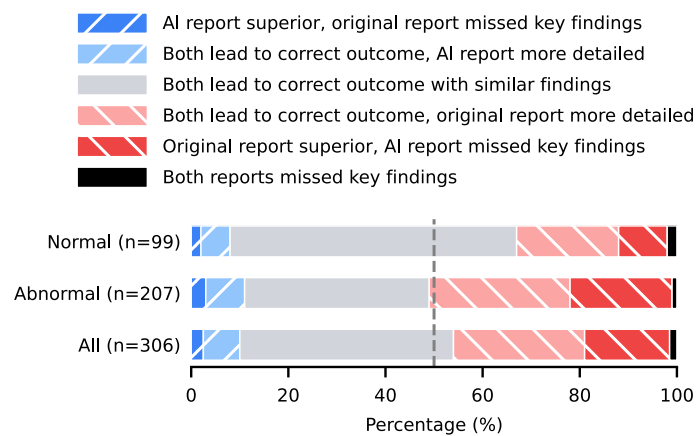
Extended data is available for this paper at <https://doi.org/10.1038/s41591-026-04626-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-026-04626-w>.

Correspondence and requests for materials should be addressed to Andrew Sellergren, David F. Steiner, Daniel Golden or Lin Yang.

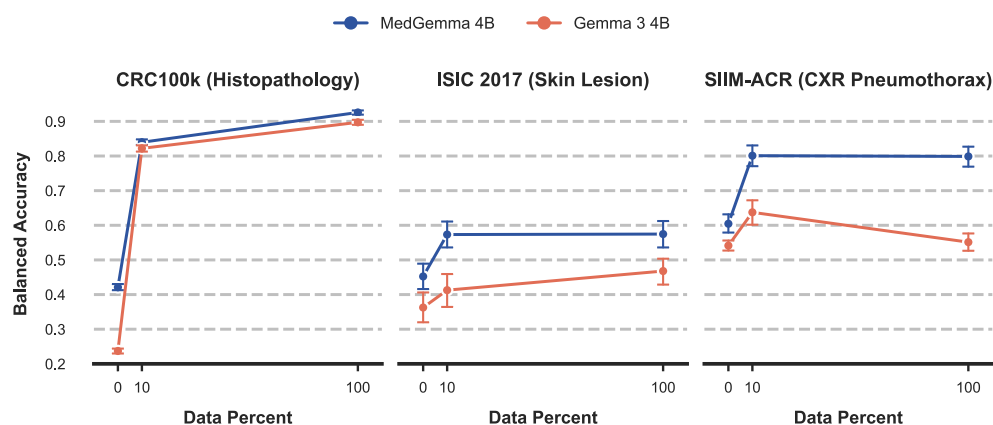
Peer review information *Nature Medicine* thanks Pranav Rajpurkar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Mattia Andreoletti and Joao Monteiro, in collaboration with the *Nature Medicine* team.

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

**Extended Data Fig. 1 | Human evaluation of chest X-ray report generation.**

The MedGemma 4B pretrained model was used to generate reports for a sample of images from the MIMIC-CXR test set, including images for which the original report identified abnormalities ('abnormal') and those for which the original report did not identify abnormalities ('normal'). A single US board-certified thoracic radiologist reviewed the MIMIC-CXR report, the AI-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. The radiologist was aware of which reports were AI generated. Across all cases, the radiologist rated 81% of MedGemma AI-generated CXR reports as likely to result in the same or superior clinical patient management when compared with the original reports.



Extended Data Fig. 2 | Fine-tuning performance of MedGemma 4B and Gemma 3 4B. Plots represent mean balanced accuracy $\pm$ 95% confidence intervals on independent test splits across varying training data fractions. To account for class imbalance, balanced accuracy was used and calculated as the unweighted mean of class-specific recall. A training fraction of 0% indicates the baseline model before fine-tuning. The unit of study was an individual image or tissue

patch for histopathology. Sample sizes are: CRC100k ($n = 7,180$ patches), ISIC 2017 ($n = 600$ dermoscopic images) and SIIM ACR ($n = 1,068$ chest X-rays). Confidence intervals were derived via non-parametric bootstrap resampling with 1,000 iterations across independent test samples; no technical replicates or repeated measurements were used.

Extended Data Table 1 | Overview of MedGemma training datasets

Modality	Dataset	No. examples*	Training stages	Description
Text-only	MedQA	9,275	Distill, RL	USMLE 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
	EHRQA [†]	10,437	RL	Synthetic long context EHR records in text form
	Synthetic	200,000	Distill	Generated Medical QA for distillation
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
	Chest ImaGenome [†]	39,968	RL	Chest X-ray images & anatomical location bounding boxes
	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	Retina images & labels
General Medical	PMC (Open Access)	41,853	Vision, PT	Single panel medical images & caption pairs

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

* Number of examples represents the available data, the final data mixture and ratios are described in the Methods section and were selected based on tuning evaluations.

[†] Only used in MedGemma 27B.

Extended Data Table 2 | Medical VQA results for SLAKE and VQA-RAD

Model	SLAKE (English)			VQA-RAD Test split from [9, 22]	
	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
BiomedGPT-B [24] [§]	85.2	-	89.9	*	*
LLaVA-Med (BioMedCLIP) [23] [§]	-	87.1	86.8	*	*
Large Models					
Med-Gemini [9]	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 this split.
- Results for this metric not available in associated report.
- § Prior reported results.

Bold values indicate the best results in each section. Note that the original VQA-RAD test set includes images that also appear in the train and test sets and we used the splits described in refs. 9,22 to avoid this contamination. The choice of these splits precludes direct comparison with externally reported results.

Extended Data Table 3 | 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
Qwen3-32B	51.4	39.5
Large Models		
DeepSeek-R1	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 using GPT-4 patient agent as reported in [15]. All other metrics were recalculated using GPT-4o as patient agent to minimize deviation from original AgentClinic report.

Bold values indicate the best results in each section. Performance of the 4B models is not reported as those models demonstrated difficulty following system instructions.

Extended Data Table 4 | Accuracy results on general, non-medical benchmarks

Type	Benchmark	MedGemma 4B	Gemma 3 4B [§]	MedGemma 27B Text	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	-	64.9

§ Prior reported results
- Not applicable

Bold values indicate the best results in each section. Note that multimodal results are not applicable to MedGemma 27B Text.

Extended Data Table 5 | Examples from the broader research community of research performed on and with MedGemma since its initial release

Reference	Modality	Task	Summary
ReXVQA [63]	CXR	Chest X-Ray VQA, including localization	MedGemma performance of 83.84, outperforming all other models evaluated and versus best radiology resident (77.27).
BRIDGE [64]	Text only Leaderboard (Multiple tasks)	87 tasks sourced from real-world clinical data	MedGemma (27B) outperformed all open LLMs in 5-shot setting, including DeepSeek-R1 (671B).
RadGame [65]	CXR	Localization	MedGemma addressed the correct anatomical location in 88%, correctly identified visual features in 96%, and specified the correct condition present in 90%.
MedThink-Bench [66]	Text only (Multiple tasks)	Assessment of LLM medical reasoning	MedGemma (27B) is the top performer, scoring 0.769 and outperforming larger commercial counterparts.
AMRG [67]	Mammography	Report generation	LoRA-tuned MedGemma (4B) outperforms the next-best model (Qwen2.5-VL-7B), particularly in METEOR and CIDEr scores, and for BI-RADS.
Kvasir-VQA-x1 [68]	Endoscopy frames	Endoscopy VQA	On this OOD task, MedGemma (4B) fine-tuned "performs nearly on par" with fine-tuning of a larger 7B generalist Qwen model.
MedVAL [69]	Text only LLM-as-Judge	MedVal-Bench: LM as a judge	Among zero-shot baselines, MedGemma (27B) achieves the highest overall performance (48.2%), slightly outperforming Llama-3.3-70B and Gemma3-27B.
AskCPG [34]	Retina, dermatology, CXR	Classification	Fine-tuning increases accuracy and F1 scores over baseline model for all modalities.
MedFM-Robust [70]	Medical imaging with image perturbation	VQA, captioning and visual grounding	MedGemma shows strong baseline performance and is the most robust tested model to image perturbation

Reporting Summary

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

Statistics

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

n/a Confirmed

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

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

Software and code

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

Data collection

N/A

Data analysis

Python and common python packages were used for data analysis (e.g. numpy, pandas, skimage, scipy, PIL)

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

Data

Policy information about [availability of data](#)

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

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

De-identified data from public and private datasets were used. Please see dataset Tables (Tables 1 and Extended Table 1) for training and evaluation, respectively.

MedGemma collection, including tutorials and instructions for downloading the model weights, can be found at <https://goo.gle/medgemma>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender N/A

Reporting on race, ethnicity, or other socially relevant groupings N/A

Population characteristics N/A

Recruitment N/A

Ethics oversight N/A

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size Benchmarks in Extended Table 1 were used for evaluation. Sample size were included in the Table. We used all samples available that's provided in these benchmarks.

Data exclusions No data exclusion

Replication Benchmark results have been rerun multiple times, consistent results are observed.

Randomization Not applicable

Blinding Not applicable

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

N/A

Novel plant genotypes

N/A

Authentication

N/A