Multimodal Large Language Models are Generalist Medical Image Interpreters

Tianyu Han [1,*], **Lisa C. Adams** [2,3,*], **Sven Nebelung** [1], **Jakob Nikolas Kather** [4,5], **Keno K. Bressem** [6,*], and **Daniel Truhn** [1,*]

1. Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Germany
2. Department of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany
3. Department of Radiology, Stanford University, Stanford, CA, United States
4. Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany
5. Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
6. Department of Radiology and Nuclear Medicine, German Heart Center Munich, Munich, Germany

* Contributed equally

Corresponding author's contact information:
Daniel Truhn, MD, PhD
University Hospital Aachen
Pauwelsstr. 30, 52074 Aachen
email: daniel.truhn@gmail.com
telephone: +491703215806

Abstract

Advanced multimodal large language models (LLM), such as GPT-4V(ision) and Gemini Ultra, have shown promising results in the diagnosis of complex pathological conditions. This raises questions about their knowledge base: Do these models deeply understand medical cases, including images, or do they simply recognize superficial patterns from extensive pre-training? We aimed to determine whether LLMs can develop useable internal representations of images, and if these representations improve the classification of medical images. We rigorously tested the performance of the open-source Flamingo-80B model, which is not specifically tailored for medical tasks, against traditional pre-training methods. The tests covered eight distinct image classification tasks in pathology, dermatology, ophthalmology, and radiology, using CLIP, Flamingo-80B, and 9B multimodal models. These tasks ranged from tissue and nuclear classification in histopathology to lesion detection in dermatology and disease grading in radiology. We systematically evaluated the model's internal image representations to determine their relevance and usefulness in medical diagnosis. Our analysis showed that the internal representation of these images in the largest model, Flamingo-80B, was more accurate in classifying medical images than in all other methods. These results held even when the number of samples available for training was small. Our results show that multimodal LLMs acquire structured knowledge in medical domains. This suggests that these models are evolving from mere pattern recognition tools into entities with broader medical generalist capabilities. This evolution underscores the potential for these models to make contributions to medical diagnosis and research, although it is important to continue to evaluate their capabilities and limitations in real-world medical settings.

Introduction

Recent advances in natural language processing have notably enhanced the capabilities of multimodal large language models (LLMs) so that they are now able to answer complex medical questions almost on par with human experts[1–4](**Figure 1**). Multimodal LLMs, being trained on vast amounts of written text[5–7], exhibit new capabilities, previously attributed only to humans like the ability to reason and to abstract away from specific problems enabling them to apply their knowledge to new, unseen problems[8]. This ability is especially important in medical training, where knowledge is largely disseminated through language, even for specialties that focus intensively on visual patterns such as radiology, pathology, or ophthalmology[9]. Provided only with a textual description of pathological image changes, medical doctors can generalize well from textual descriptions to image interpretation. Replicating this ability using deep learning (DL) models has been a long-sought goal of research[10–17].

To enhance performance in medical tasks, specialized LLMs have been developed, primarily through augmenting generalist models with extensive training on medical data[18–21].

However, recent literature questions the need to transform generalist models into specialists. Using advanced prompting techniques, generalist models can outperform their specialist counterparts on medical tasks[22]. The success of these prompting techniques in medical contexts indicates that these models inherently contain accurate representations of domain-specific knowledge. The objective of our study was to investigate the extent of generalist models' comprehension of medical data, with a specific emphasis on medical imaging.

In particular, our approach employs open-source multimodal LLM models, notably those not specifically tailored to medical tasks. We focus on images from four medical fields heavily reliant on image classification: pathology, dermatology, ophthalmology, and radiology. For each field, we selected two use cases and applied them to distinct datasets, observing how the LLMs internally represented these images and whether this representation could distinguish between various medical subclasses.

Our findings indicate that the general natural language pre-training undergone by LLMs may offer advantages over more specialized, task-specific pre-training methods in certain medical contexts. This comparison includes benchmarks in image- and language-pretraining, such as those reported in the recent work by Huang et al[23,24]. While this suggests a promising direction for the application of LLMs in medical image analysis, it also highlights the need for further research and validation[25–27]. Our study aims to contribute to the ongoing dialogue on the utility of LLMs in medical science, particularly in integrating and interpreting complex visual and textual data - a prerequisite for foundational models[28–30].

Methods

Ethics Approval

This study was conducted in accordance with the tenets of the Declaration of Helsinki and was approved by the local institutional review board (EK259/22).

Patient Cohorts and Imaging Data

In this study, we systematically examined medical imaging datasets across four key medical disciplines: pathology, dermatology, ophthalmology, and radiology. We conducted two specific image classification tasks within each discipline, resulting in a total of eight distinct tasks (**T**), see **Figure 2a** and **Table 1**:

Tissue Classification in Histopathology Images (T1): Using the NCT-CRC-HE-100K dataset, this task includes histological imaging data from 136 colorectal cancer patients. Following the dataset partitioning proposed by Kather et al[31], we formed a training set of 100,000 image patches from 86 patients and a test set of 7,180 patches from 50 patients. Each patch, measuring 224×224 pixels, is classified into one of nine tissue categories: adipose tissue, background, debris, lymphocytes, mucus, smooth muscle, normal colonic mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium[31].

Nuclear Classification in Histopathology Images (T2): This task uses the PanNuke dataset, which contains 7,558 pan-cancer images from 19 different organ types[32]. These images, which were annotated by Gamper et al., include various nuclear categories such as neoplastic, inflammatory, connective, epithelial, and dead tissue, including both apoptotic and necrotic cells.

Lesion Detection in Dermatology (T3): For this task, we utilized the 2018 International Skin Imaging Collaboration (ISIC) Challenge dataset, comprising 10,208 training and 1,512 testing images of various skin lesions. Classifications include melanoma, basal cell carcinoma, and several other lesion types, as detailed in the work by Tschandl et al[33,34].

Melanoma Classification in Dermatology (T4): Derived from the ISIC 2020 Challenge, this task includes dermatology data with images labeled as benign or malignant[35]. The dataset, which differs from the 2018 challenge, includes 26,045 images for training and 7,081 for testing, stratified by patient (1,644 patients for training, 412 for testing).

Diabetic Retinopathy Grading in Fundoscopic Images (T5):: We sourced data from the 2015 EyePACS Diabetic Retinopathy Detection Challenge[36] and the APTOS-2019 Blindness Detection Challenge[37], totaling 88,700 fundoscopies from 44,350 patients. The combined dataset was divided into 73,622 training images (only EyePACS) and 18,740 testing images (from EyePACS (7,539 patients) and APTOS-2019).

Glaucoma Detection in Fundoscopic Images (T6): This task incorporates data from the AIROGS[38] and ODIR-2019[39] challenges, resulting in a large dataset of 101,442 fundoscopies from 54,274 patients for training and 7,000 fundoscopies from 3,500 patients for testing.

Lung Disease Detection in Chest Radiographs in Radiology (T7): Using the 'PadChest' cohort, this task focuses on radiology data with 86,715 chest radiographs from 59,975 patients for training and 7,943 radiographs from 7,272 patients for testing[40,41]. The dataset includes 174 radiographic findings and 19 radiological diagnoses[41].

Osteoarthritis Grading in Knee Radiographs in Radiology (T8): Employing data from the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST), this task involves grading osteoarthritis in knee radiographs[42,43]. Following the methodology of Han et al.[15], we constructed a dataset with 56,185 training images from 6,425 patients and 9,904 testing images from 1,095 patients.

Table 1: Details on eight image classification tasks.

Task	Type	Datasets	Training Set		Testing Set	
			Images	Patients	Images	Patients
T1	Histopathology	NCT-CRC-HE-100K[31]	100,000	136	7,180	50
		CRC-VAL-HE-7K[31]				
T2	Histopathology	PanNuke[32]	4,971	N/A	1,263	N/A
T3	Dermatology	HAM10000[34]	10,208	N/A	1,512	N/A
		ISIC 2018[33]				
T4	Dermatology	SIIM-ISIC 2020[35]	26,045	1,644	7,081	412
T5	Ophthalmology	EyePACS[36]	73,622	36,811	15,078	7,539
		APTOS-2019[37]	0	0	3,662	N/A
T6	Ophthalmology	AIROGS[38]	101,442	54,274	0	0
		ODIR-2019[39]	0	0	7,000	3,500
T7	Radiology	PadChest[41]	86,715	59,975	7,943	7,272
T8	Radiology	OAI[42,43]	39,921	3,831	7,108	677
		MOST[42,43]	16,264	2,594	2,796	418

NEJM Image Challenge Benchmarking

In this study, we collected 931 clinical cases from the NEJM Image Challenge from October 2005 to August 2023. Each case presented a medical image accompanied by a short text describing the clinical context, culminating in a specific question such as "What is the diagnosis?" (see **Figure S1** for an example). We provided five possible answers for each case and tasked DeepMind's Flamingo model with selecting the correct answer.[6] The dataset covered a wide range of medical fields, including pathology, dermatology, ophthalmology, and radiology, providing a comprehensive mix of medical imaging data. Statistics on the number of correct answers provided by NEJM readers were used to stratify the difficulty of the questions into five equal intervals according to the percentage of correct answers provided by human readers.[3]

We used a few-shot, in-context learning approach to test Flamingo on the NEJM cases.[44] This involved using the first two cases from the dataset (dated October 13th and 20th, 2005) as

initial examples for the model (**Figure S2**). The remaining 929 cases were then used as a test set to assess the model's ability to interpret medical images across different disciplines.

Multimodal LLMs

We used the open-source Flamingo architecture, [45] which was trained by Hugging Face M4 and is available in two sizes: Flamingo-80B with 80 billion parameters and Flamingo-9B with 9 billion parameters. Both models are VLMs that accept text interleaved with images and output free-form text. Flamingo combines a pre-trained LLM (LLaMA-65B for Flamingo-80B and Llama-7B for Flamingo-9B[46]) and a pre-trained Vision Transformer (ViT, 632M parameters[47]) via a transformer-based mapper (Perceiver Sampler[48]). To fuse vision and text signals, Flamingo uses cross-attention layers interleaved with LLM residual blocks (see **Figure 2c**). LLaMA-65B was pre-trained on 1.4 trillion tokens from publicly available data sources, including Wikipedia, arXiv, Github, Books, StackExchange, C4, and CommonCrawl[46]. The ViT was pre-trained on 2.3 billion images obtained from the web as part of the LAION-5B dataset[49]. The combined Flamingo model was then further pre-trained for its perceiver samplers and cross-attention blocks on 141 million interleaved image-text documents and 353 million images[45].

Testing the Models' Medical Image Interpretation

To test the medical reasoning of the models and their ability to stratify medical images for downstream tasks, we use a method similar to recently published approaches [50–53], i.e., we present the respective images to the model along with a general prompt, e.g., "What can you see on this radiological image?". We then extract the representation of the images in the model's internal latent space and test whether these representations can be used for classification by a simple linear logistic regression model, see **Figure 2c**. This concept is called "probing the model" and tests whether the internal representation of the images is linearly separable, i.e. whether the LLM has allocated healthy and pathological images to separate regions of its high-dimensional space.

CLIP as a Comparison Model

We used OpenAI's CLIP (Contrastive Language-Image Pre-training) as a benchmark to evaluate Flamingo's performance. CLIP, specifically the CLIP-ViT-B/32 model, is trained on a corpus of over 400 million Internet-sourced image-text pairs, providing robust "zero-shot" learning capabilities[54]. We use this baseline model in all tasks T1-T8. As a second baseline model, focused only on the pathology tasks, we employ PLIP (Pathology Language-Image Pre-training), which has been trained with contrastive learning specifically on pathology images sourced from X (formerly Twitter) and has recently been presented as a foundational model with state-of-the-art performance in histopathology[23].

Image Pre-processing

Images larger than 1024×1024 pixels were downsampled to 1024×1024 pixels and underwent normalization relative to their maximum pixel value to ensure uniformity across the datasets. T2 and T8 required specific preprocessing: in T2, images of nuclei were processed according to the work of Huang et al[23]. The image was considered 'malignant' if the total number of neoplastic cells was more than ten and covered more than 30% of the total cells. Images were

considered 'benign' if no neoplastic cells were present. This resulted in 2,866 malignant images and 3,368 benign images. For T8, knee radiographs were preprocessed to include only a 140 mm×140 mm region using a pre-trained hourglass network reported by Tiulpin et al.[55]

Computational Resources

We use four NVIDIA A6000 (48GB) GPUs on a local server system to probe the models. To train the logistic regression model on the internal probes of Flamingo activations, an NVIDIA RTX 3090 (24GB) GPU was used.

Evaluation and Statistical Analysis

For T3 to T8, the performance of the classifiers was evaluated by the area under the receiver-operator curve (AUC). For T1 and T2, the classification performance was evaluated by the F1 score according to Huang et al.[23] Standard deviations (SDs) and P values were calculated using bootstrapping with 1,000 replicates and paired 2-tailed t-tests.

Results

We present our results as follows: First, we show Flamingo's performance on the NEJM Image Challenge dataset by prompting it with medical speech-image questions and recording the output as text (**Figure 1**). This mimics direct human interaction with the model. We then use T1 through T8 to explore the model's internal medical reasoning capabilities and compare its performance to CLIP on large datasets across eight application cases (**Figure 3-5**). Finally, we show that the internal image representation allows for highly data-efficient development of AI models when limited labels are available, achieving state-of-the-art performance with a fraction of the data of other models (**Figure 6**).

Accuracy in a Complex Diagnostic Challenge

When analyzing 929 diagnostic cases, Flamingo-80B's primary diagnosis matched the final diagnosis in 40.4% (375 of 929) of cases (**Figure 1d**). When the model was prompted three times in succession, it included the correct diagnosis in 54.3% (504 of 929) of cases, as determined by stochastic top-K sampling with T=1.0 and top k=50. Notably, Flamingo-80B's performance outperformed guesswork at various levels of difficulty, except for the most difficult category (**Figure 1d**). In **Figure 1a-c**, we illustrate selected Flamingo-80B responses and their rationale. These results highlight Flamingo-80B's ability to provide medical insight and to integrate medical knowledge, albeit with the need for careful interpretation and validation in real-world settings.

Systematic Investigation

To determine whether the ability of multimodal LLM models to answer complex medical questions stems from an understanding of medical principles, we presented image data with textual prompts to Flamingo-80B and Flamingo-9B, as well as using OpenAI's CLIP as a benchmark. Our focus was on analyzing the internal state representations of these models to determine their medical relevance.

Classification in Pathology

The colorectal tissue classification task (T1) focused on classifying tissue into nine categories based on hematoxylin & eosin (H&E)-stained histologic images from a human colorectal cancer (CRC) cohort. In this task, a linear classifier was trained on internal activations obtained from multimodal LLMs and the CLIP model, analyzing a total of 7,158 histopathological image patches. The results showed that Flamingo-80B's internal representations achieved a higher average F1 score of 0.892 as compared to the CLIP method, which scored 0.764. Notably, Flamingo-80B also outperformed the visual language foundation model developed by Huang et al.[23], which was pre-trained on Twitter for domain-specific data, with an F1 score of 0.892 versus 0.877. Detailed results for the different categories can be found in **Figure 3a-i**.

In the nuclear classification task (T2), our goal was to discriminate between benign and malignant cases among samples from 19 different organs using the PanNuke dataset (**Figure S6**). By applying a linear classifier to the internal activations derived from both multimodal LLMs and the CLIP model, Flamingo-80B demonstrated superior performance. Specifically, its internal representations yielded a consistently higher F1 score of 0.870 (95% CI: [0.847 to 0.891]) compared to the baseline CLIP method's 0.797 (95% CI: [0.774 to 0.821]) (t-

statistic=139.7, P<0.001), as detailed in **Figure 3j**. These results collectively confirm the advanced capabilities of multimodal LLMs over traditional pre-training methods in histopathology, even matching the accuracies of specialized foundation models that rely on domain-specific data.

Classification in Dermatology

The skin lesion detection task (T3) involves the multiclass classification of dermatological images into seven classes: melanoma, basal cell carcinoma, actinic keratosis carcinoma, melanocytic nevus, benign keratinocytic lesions, dermatofibroma, and vascular lesions. After training the linear classifier on the internal activations extracted from multimodal LLMs and the CLIP model Flamingo-80B's internal representations resulted in a consistently higher AUC as compared with the baseline CLIP method in all seven classes, see **Figure 3k-q** for a more detailed breakdown (P<0.001 for all).

The second skin lesion classification task (T4) on a separate dataset classified 33,126 dermatological images into malignant or benign lesions. Following the same architecture as above, Flamingo-80B achieved a significantly higher AUC on this task than CLIP (0.885, 95% CI: [0.859 to 0.909] vs. 0.834, 95% CI: [0.810 to 0.857], P<0.001), see **Figure 3r**.

Classification in Ophthalmology

T5 focuses on the detection of diabetic retinopathy using over 90,000 fundus photographs in the US and India. Flamingo-80B shows superior performance in grading diabetic retinopathy (see **Figure 4**), especially in detecting proliferative and severe diabetic retinopathy (**Figure 4a, b**), achieving state-of-the-art results (AUC=0. 949, 95% CI: 0.939 to 0.958; and AUC=0.903, 95% CI: 0.889 to 0.917) and significantly outperformed the baseline CLIP model (AUC=0.883, 95% CI: 0.870 to 0.896 and AUC=0.826, 95% CI: 0.808 to 0.846; P< 0.001 for both classes). Performance in detecting mild diabetic retinopathy is lower for all three models (Figure 4d), possibly due to class imbalance and labeling ambiguity, with Flamingo-80B performing best with an AUC of 0.629 (95% CI: 0.612 to 0.644).

T6 addresses another significant visual impairment cause, glaucoma, assessed in a large patient cohort from Beijing, China, comprising 3,500 individuals[56]. Here again, the probe trained on the Flamingo-80B activations showed superior performance in AUC (0.868) compared to both its smaller variant, Flamingo-9B (AUC: 0.843; P<0.001), and the baseline CLIP model (AUC: 0.716; P<0.00, **Figure 4f**).

Classification in Radiology

The chest X-ray classification task (T7) aims at allocating 54 radiographic findings to chest X-rays from the PadChest dataset. We utilized 94,658 chest X-rays of which 27.9% were labeled manually by board-certified radiologists. A subset of 7,943 manually labeled chest X-rays was set aside for testing. After training the linear classifier on the internal activations of the multimodal LLMs, Flamingo-80B led to an AUC of at least 0.90 in 7 findings and of at least 0.70 in 40 findings. CLIP achieved these AUC thresholds in none and only 6 findings, respectively, see **Figure 5**.

T8 investigates the performance of diagnosing osteoarthritis (OA) in knee X-rays. OA was graded based on manual labels by board-certified radiologists.[15] Again training a linear model

on the internal activations led to the superior performance of Flamingo-80B in severe OA (0.971, 95% CI: 0.965 to 0.976), moderate OA (0.870, 95% CI: 0.860 to 0.880), and no OA (0.815, 95% CI: 0.807 to 0.824). CLIP's performance was consistently lower with an AUC of (0.907, 95% CI: 0.894 to 0.920) in severe OA, (0.734, 95% CI: 0.720 to 0.748) in moderate OA, and (0.706, 95% CI: 0.696 to 0.715) in no OA, see **Figure 4g-k**.

Multimodal LLMs are data efficient

Our goal was to determine whether LLMs' inherent knowledge and inference capabilities could facilitate the development of AI models using a reduced number of labels. To this end, we conducted a series of label efficiency experiments. These experiments were designed to determine the minimum amount of training data and labels required for LLMs to achieve specific performance benchmarks on various medical tasks.[29]

Our results were particularly striking with Flamingo-80B. Using only 10% of the training data, Flamingo-80B was able to retain good performance across four medical disciplines. Specifically, it maintained 95.8% (comparing an F1 score of 0.855 with 10% data to an F1 score of 0.892 with 100% data), 94.3% (comparing an AUC of 0.892 with 10% data to an AUC of 0.945 with 100% data), 95. 2% (comparing an AUC of 0.764 with 10% data to an AUC of 0.803 with 100% data) and 94.7% (comparing an AUC of 0.767 with 10% data to an AUC of 0.810 with 100% data) of its peak performance in pathology, dermatology, ophthalmology, and radiology, respectively. Detailed results of these findings are shown in **Figure 6**.

These results suggest that the knowledge and inference capabilities embedded in multimodal LLMs are highly effective, enabling the development of AI models with minimal labeled data.[28] This feature of LLMs holds great promise for applications where large labeled datasets are not readily available.

Discussion

In our study, we present evidence that generalist models such as Flamingo-80B can inherently understand medical images and, in some cases, even outperform specialized models, such as PLIP[23], to achieve new state-of-the-art performance. Using representations for generalist models may thus offer a data-effective solution for developing classification models in the medical domain.

In the past, the predominant technique for solving tasks in particular domains such as medicine was the training of specialist models. This led to the creation of first-generation specialized language models such as PubMedBERT[57] and BioGPT[58], and multiple other models, specialized for electronic health records[59] or diagnostic applications in radiology[60] or ophthalmology[29]. The most advanced medical language model is the proprietary model Med-PaLM 2[18,61], a 340 billion parameter model from Google, fine-tuned from Palm 2[62].
However, specialist models now seem to be losing their advantage over generalist models. Today, the best-performing model on various benchmarks is the generalist GPT-4[8,22], raising the question of whether fine-tuning is still needed or whether generalists will soon be able to solve all tasks, making specialist models obsolete. For example, GPT-4 with specialized prompting achieves an accuracy of 90.2% on the MedQA[63] benchmark of USMLE-style questions beating Med-PaLM 2 which achieved 86.5%[22].

However, comparing models to GPT-4 is inherently flawed because not much is known about this proprietary model by OpenAI, such as model size, architecture, and amount of training data[5]. It is conceivable that GPT-4's training dataset encompasses an extensive range of biomedical knowledge, possibly more data than specialized models were trained on[5], which expresses a strong performance on most specialized benchmarks. Furthermore, language models benefit immensely from scale[64,65], and the size of GPT-4, although unknown, is likely an order of magnitude larger than that of other models. This may explain why this generalist model, with appropriate prompting techniques, excels in several specialized domains such as medicine.

Our research differs by focusing on the open-source VLM Flamingo-80B, ensuring a more equitable comparison. We show that Flamingo-80B, a generalist VLM, inherently possesses medical knowledge and excels at classification tasks without specialized training. We performed an extensive evaluation of eight datasets from four medical specialties comprising more than 450,000 medical images and demonstrated the wide applicability of our findings. We thus conclude that VLMs encode general medical knowledge and are suitable as generalist medical image interpreters. This finding suggests a reevaluation of the current approach to AI in medicine, where specialist models are trained for new applications, and argues for a more integrated use of generalist models in the field. Generalist VLMs offer a versatile, cost- and data-efficient alternative to the development of multiple specialized models. We demonstrated that Flamingo-80B allows for the creation of highly performant image classification models based on the internal representations of the model, using only 10% of the training data. Given the general sparsity of medical training data and the high costs of labeling data with domain experts, the use of models such as Flamingo-80B possesses great potential. In addition, their inherent knowledge and ability to process information from other domains can facilitate the linking of different domains within the medical field and the incorporation of existing knowledge[18,26].

Limitations

Our work has limitations and leaves room for future research. Specifically, we performed a proof-of-concept and focused solely on imaging information. Therefore, we did not investigate the fusion of imaging information with more complex textual information, such as patient reports or patient history. Additionally, the model exhibited hallucinations when answering some of the clinical vignette questions for the NEJM challenge. We provided examples in **Figure S3** and **Figure S4** but did not conduct a thorough analysis of hallucinated findings. A third limitation is that the NEJM challenge questions are not a factual representation of the clinical workflow, but rather a vignette of clinical cases used to evaluate the LLM's clinical reasoning skills. Follow-up studies are necessary to establish the real clinical use of such models. Most importantly, we used LLaMA as the LLM backbone. While there are more powerful proprietary models like GPT4V by OpenAI and Gemini Ultra by Google, LLAMA is the current state-of-the-art among open-source models. We were unable to test these proprietary models due to their closed nature, but we anticipate that they, along with future open-source LLMs, will result in even more high-performing vision-language models.

Conclusions

The development of large generalist visual language models, such as Flamingo-80B is transforming medical diagnostics. The performance of Flamingo-80B, particularly its ability to create high-performing image classification models using substantially less training data, highlights the model's innate medical knowledge and its applicability in scenarios characterized by data scarcity and high costs of expert data labeling. This efficiency in leveraging internal representations of medical imagery opens new possibilities for medical AI, particularly in domains where data is limited.

Data availability

The NEJM challenge questions are available to the public via: https://www.nejm.org/image-challenge. The validation datasets are publicly available and can be accessed from the following: Kather Colon (https://zenodo.org/record/1214456); PanNuke (https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke); ISIC-2018 (https://challenge.isic-archive.com/data/#2018); ISIC-2020 (https://challenge.isic-archive.com/data/#2020); EyePACS Diabetic Retinopathy Detection (https://www.kaggle.com/c/diabetic-retinopathy-detection/); APTOS-2019(https://www.kaggle.com/c/aptos2019-blindness-detection); AIROGS (https://zenodo.org/records/5793241); ODIR-2019 (https://odir2019.grand-challenge.org/Download/); PadChest (https://bimcv.cipf.es/bimcv-projects/padchest/); OAI (https://nda.nih.gov/oai/query-download); MOST (https://most.ucsf.edu/multicenter-osteoarthritis-study-most-public-data-sharing).

Code availability

The source codes can be accessed at https://github.com/peterhan91/Multimodal-Probes. The weights of open-sourced Flamingo models can be downloaded via https://huggingface.co/HuggingFaceM4/idefics-80b-instruct and https://huggingface.co/HuggingFaceM4/idefics-9b-instruct.

OpenAI's CLIP model can be downloaded via https://huggingface.co/openai/clip-vit-base-patch32.
Inferencing of multimodal LLMs was performed using Huggingface transformers library (v.4.34.0.dev0, https://huggingface.co/docs/transformers/index) and PyTorch (v.2.0.1, https://pytorch.org/). Analysis of LLM's representations was performed using Python (v.3.9.17, https://www.python.org/), scikit-learn (v.1.3.0, https://scikit-learn.org/stable/), and SciPy (v.1.11.1, https://scipy.org/).

Acknowledgments

None.

Author contributions

T.H., L.C.A., K.K.B., and D.T. devised the study concept, and D.T. performed the reader tests. T.H. wrote the code and performed the performance studies. T.H. and D.T. did the statistical analysis. T.H., L.C.A., K.K.B., and D.T. wrote the first draft of the manuscript. All authors contributed to correcting the manuscript.

Competing interests

D.T. holds shares in StratifAI GmbH and reports speaker fees from Bayer, Germany. K.K.B. reports speaker fees from Canon Medical Systems Corporation and GE HealthCare.
No other disclosures are reported.

Funding

DT is supported by the German Federal Ministry of Education and Research (SWAG, 01KD2215A; TRANSFORM LIVER), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091). K.K.B. reports grants from the European Union (101079894) and Wilhelm-Sander Foundation and serves as an advisor for the EU Horizon 2020 LifeChamps project (875329) and the EU IHI Project IMAGIO (101112053).

References

1. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Eriksen, A. V., Möller, S. & Ryg, J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* (2023) doi:10.1056/AIp2300031.
3. Han, T. *et al.* Comparative Analysis of GPT-4Vision, GPT-4 and Open Source LLMs in Clinical Diagnostic Accuracy: A Benchmark Against Human Expertise. 2023.11.03.23297957 Preprint at https://doi.org/10.1101/2023.11.03.23297957 (2023).
4. Buckley, T., Diao, J. A., Rodman, A. & Manrai, A. K. Accuracy of a Vision-Language Model on Challenging Medical Cases. Preprint at https://doi.org/10.48550/arXiv.2311.05591 (2023).
5. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
6. Alayrac, J.-B. *et al.* Flamingo: a Visual Language Model for Few-Shot Learning. Preprint at https://doi.org/10.48550/arXiv.2204.14198 (2022).
7. Driess, D. *et al.* PaLM-E: An Embodied Multimodal Language Model. Preprint at https://doi.org/10.48550/arXiv.2303.03378 (2023).
8. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at http://arxiv.org/abs/2303.13375 (2023).
9. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
10. Yim, J. *et al.* Predicting conversion to wet age-related macular degeneration using deep learning. *Nat. Med.* **26**, 892–899 (2020).
11. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
12. Barata, C. *et al.* A reinforcement learning model for AI-based decision support in skin cancer. *Nat. Med.* **29**, 1941–1946 (2023).
13. Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data

using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).

14. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
15. Han, T. *et al.* Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nat. Mach. Intell.* **4**, 1029–1039 (2022).
16. Han, T. *et al.* Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. *Nat. Commun.* **12**, 4315 (2021).
17. Han, T. *et al.* Breaking medical data sharing boundaries by using synthesized radiographs. *Sci. Adv.* **6**, eabb7973 (2020).
18. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* 1–9 (2023) doi:10.1038/s41586-023-06291-2.
19. Lu, M. Y. *et al.* A Foundational Multimodal Vision Language AI Assistant for Human Pathology. Preprint at http://arxiv.org/abs/2312.07814 (2023).
20. Moor, M. *et al.* Med-Flamingo: a Multimodal Medical Few-shot Learner. Preprint at http://arxiv.org/abs/2307.15189 (2023).
21. Han, T. *et al.* MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data. Preprint at https://doi.org/10.48550/arXiv.2304.08247 (2023).
22. Nori, H. *et al.* Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Preprint at http://arxiv.org/abs/2311.16452 (2023).
23. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
24. Lu, M. Y., Chen, B. & Mahmood, F. Harnessing medical twitter data for pathology AI. *Nat. Med.* **29**, 2181–2182 (2023).
25. Harris, E. Large Language Models Answer Medical Questions Accurately, but Can’t Match Clinicians’ Knowledge. *JAMA* **330**, 792–794 (2023).
26. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as

scientific reasoning engines, not knowledge databases. *Nat. Med.* **29**, 2983–2984 (2023).

27. Han, T. *et al.* Medical Foundation Models are Susceptible to Targeted Misinformation Attacks. Preprint at https://doi.org/10.48550/arXiv.2309.17007 (2023).
28. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
29. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
30. Bommasani, R., Hudson, D. A., Altman, E. A. R. & Arora, S. On the Opportunities and Risks of Foundation Models.
31. Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
32. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A. & Rajpoot, N. PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. in *Digital Pathology* (eds. Reyes-Aldasoro, C. C., Janowczyk, A., Veta, M., Bankhead, P. & Sirinukunwattana, K.) 11–19 (Springer International Publishing, 2019). doi:10.1007/978-3-030-23937-4_2.
33. Codella, N. *et al.* Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). Preprint at https://doi.org/10.48550/arXiv.1902.03368 (2019).
34. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
35. Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **8**, 34 (2021).
36. Diabetic Retinopathy Detection. https://kaggle.com/competitions/diabetic-retinopathy-detection.
37. APTOS 2019 Blindness Detection. https://kaggle.com/competitions/aptos2019-

blindness-detection.

38. de Vente, C. *et al.* AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge. Preprint at https://doi.org/10.48550/arXiv.2302.01738 (2023).
39. Hemelings, R. *et al.* A generalizable deep learning regression model for automated glaucoma screening from fundus images. *Npj Digit. Med.* **6**, 1–15 (2023).
40. Han, T. *et al.* Reconstruction of Patient-Specific Confounders in AI-based Radiologic Image Interpretation using Generative Pretraining. Preprint at https://doi.org/10.48550/arXiv.2309.17123 (2023).
41. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797 (2020).
42. Eckstein, F., Wirth, W. & Nevitt, M. C. Recent advances in osteoarthritis imaging—the Osteoarthritis Initiative. *Nat. Rev. Rheumatol.* **8**, 622–630 (2012).
43. Segal, N. A. *et al.* The Multicenter Osteoarthritis Study: Opportunities for Rehabilitation Research. *PM&R* **5**, 647–654 (2013).
44. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
45. Laurençon, H. *et al.* OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. Preprint at https://doi.org/10.48550/arXiv.2306.16527 (2023).
46. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at http://arxiv.org/abs/2302.13971 (2023).
47. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
48. Jaegle, A. *et al.* Perceiver: General Perception with Iterative Attention. Preprint at https://doi.org/10.48550/arXiv.2103.03206 (2021).
49. Schuhmann, C. *et al.* LAION-5B: An open large-scale dataset for training next generation image-text models. Preprint at https://doi.org/10.48550/arXiv.2210.08402 (2022).

50. Gurnee, W. & Tegmark, M. Language Models Represent Space and Time. Preprint at https://doi.org/10.48550/arXiv.2310.02207 (2023).
51. Li, K. *et al.* Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. Preprint at https://doi.org/10.48550/arXiv.2210.13382 (2023).
52. Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances. *Comput. Linguist.* **48**, 207–219 (2022).
53. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. Preprint at https://doi.org/10.48550/arXiv.1610.01644 (2018).
54. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).
55. Tiulpin, A., Melekhov, I. & Saarakkala, S. KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks. Preprint at https://doi.org/10.48550/arXiv.1907.12237 (2019).
56. ODIR-2019 - Grand Challenge. *grand-challenge.org* https://odir2019.grand-challenge.org/.
57. Gu, Y. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **3**, 2:1-2:23 (2021).
58. Luo, R. *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
59. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* 1–6 (2023) doi:10.1038/s41586-023-06160-y.
60. Smit, A. *et al.* CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. Preprint at http://arxiv.org/abs/2004.09167 (2020).
61. Singhal, K. *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2305.09617 (2023).
62. Anil, R. *et al.* PaLM 2 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2305.10403 (2023).

63. 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).
64. Rae, J. W. *et al.* Scaling Language Models: Methods, Analysis & Insights from Training Gopher. Preprint at http://arxiv.org/abs/2112.11446 (2022).
65. Wei, J. *et al.* Emergent Abilities of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2206.07682 (2022).

Figures

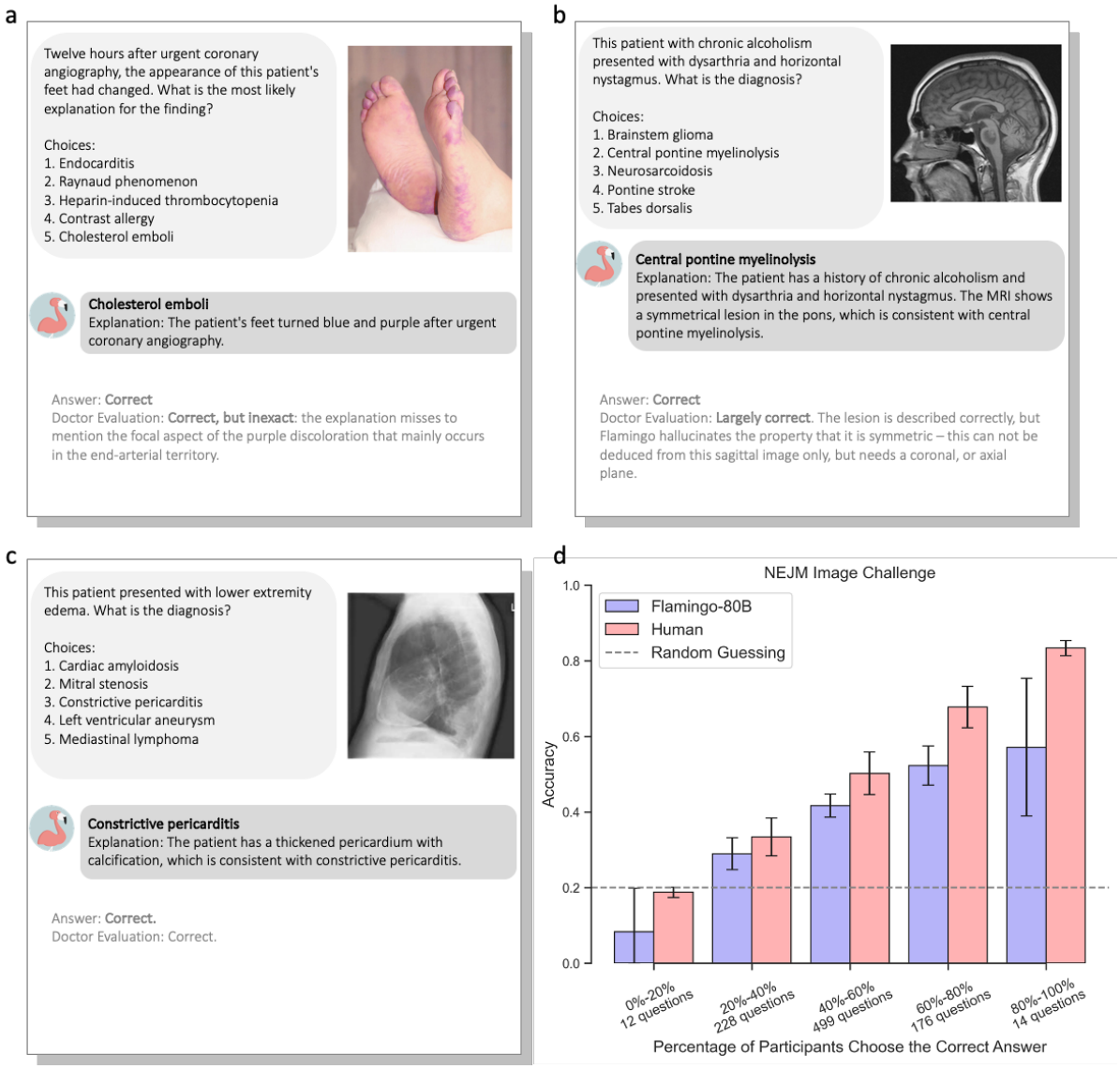


Figure 1: Performance of the multimodal LLM with 80 billion parameters on the NEJM Image Challenge Cases. (a)-(e): Selected NEJM cases correctly answered by the multimodal 80B LLM. The model provided the answer along with an explanation that was checked by a board-certified radiologist with 12 years of experience. (f): performance of Flamingo-80B in the NEJM challenge as compared to non-selective human participants. Bars indicate accuracy means; vertical lines indicate standard deviations.NEJM - The New England Journal of Medicine.

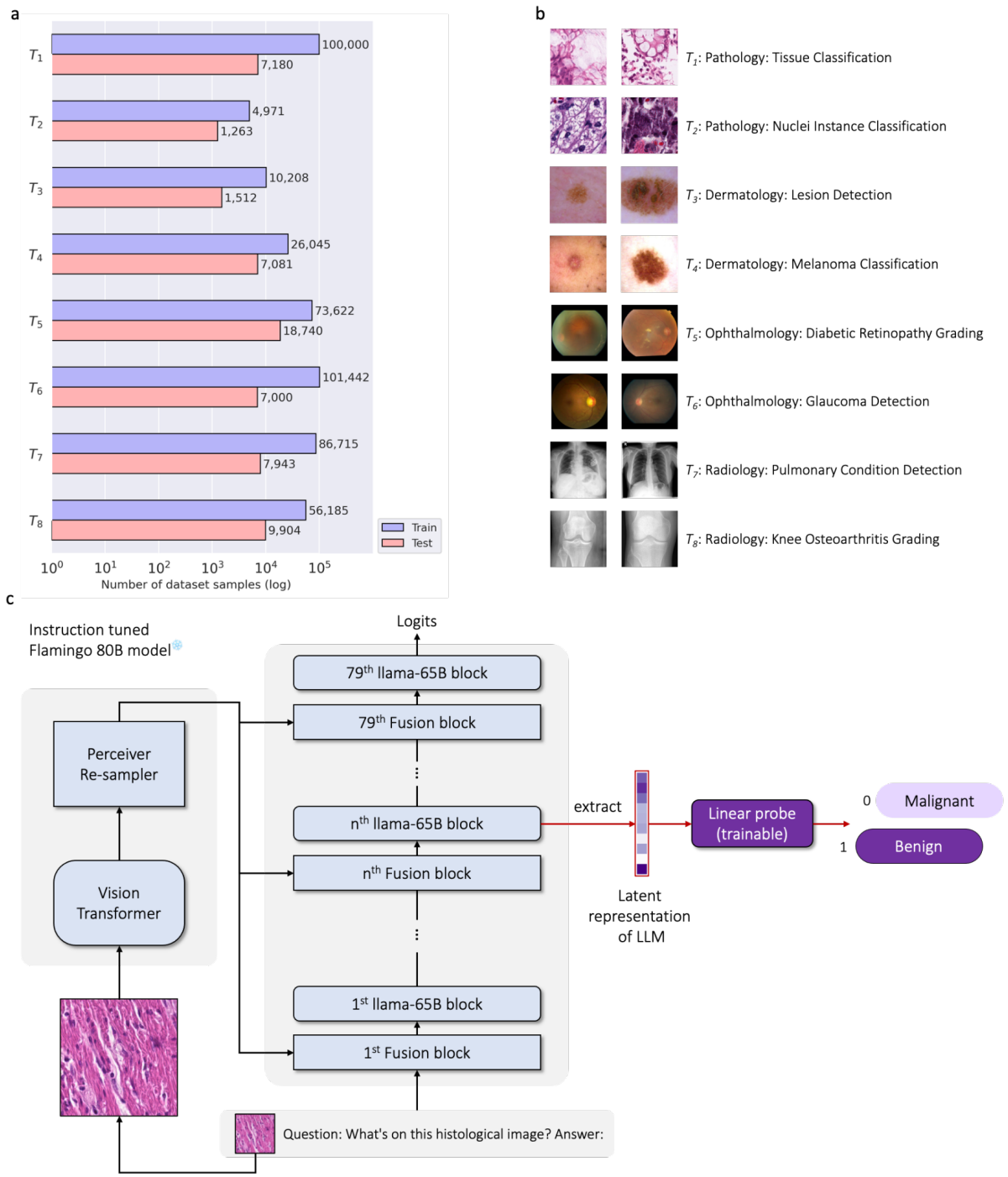
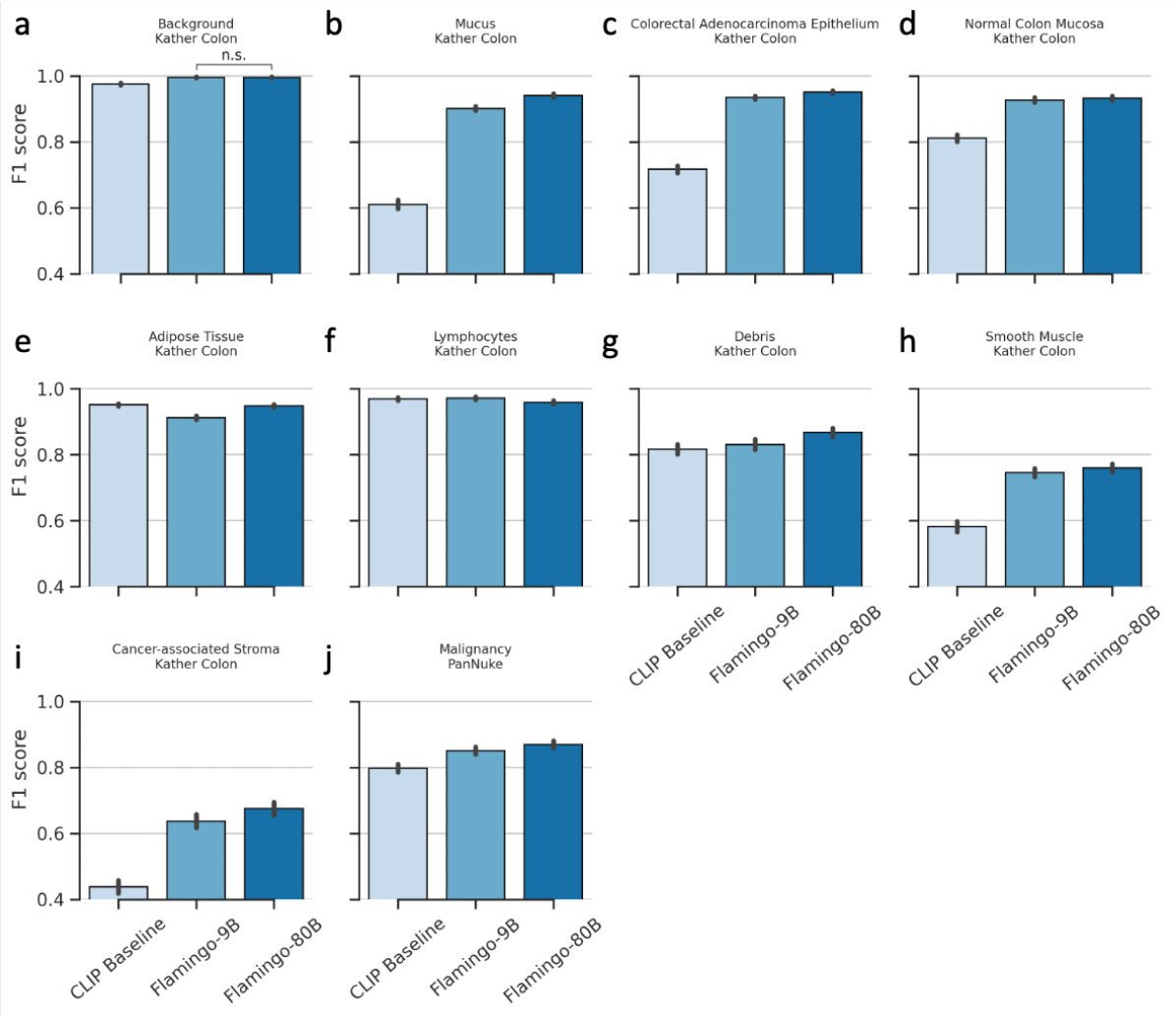


Figure 2: Setup of Experiments. (a)-(b): Flamingo (80B and 9B) models were evaluated on eight image classification tasks of four medical imaging domains. (c): Visualization of the probing of internal states used for the classification. Both vision and LLM-trained weights are frozen during probing (colored in light blue).

Pathology



Dermatology

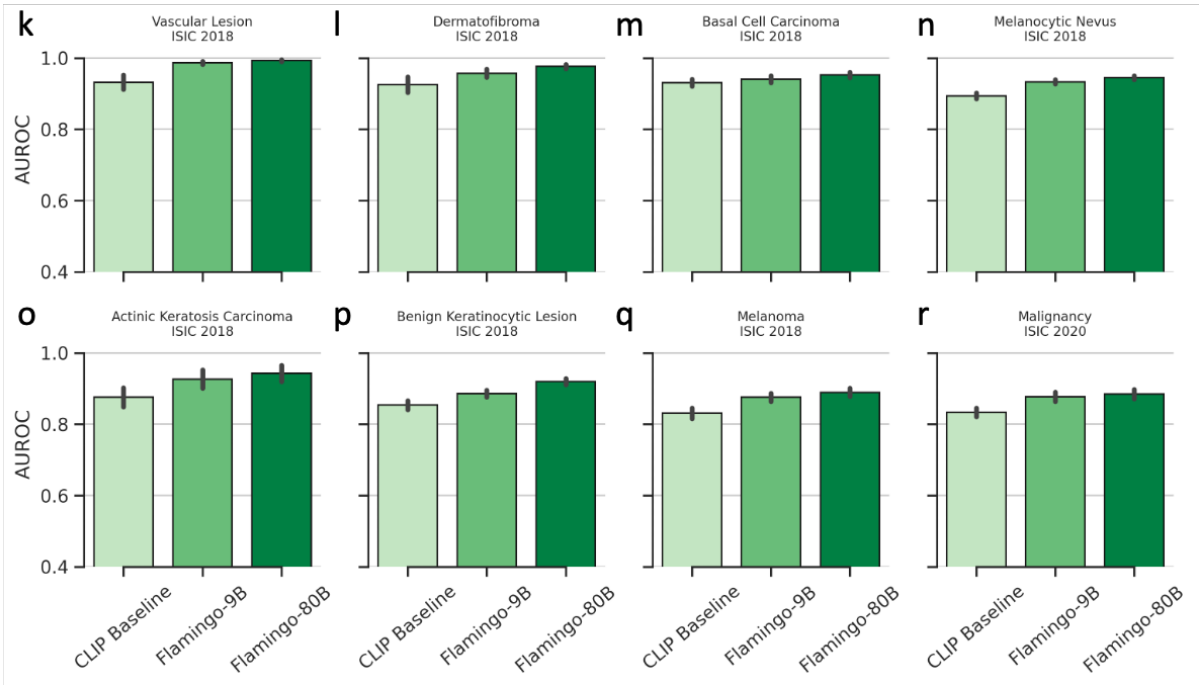
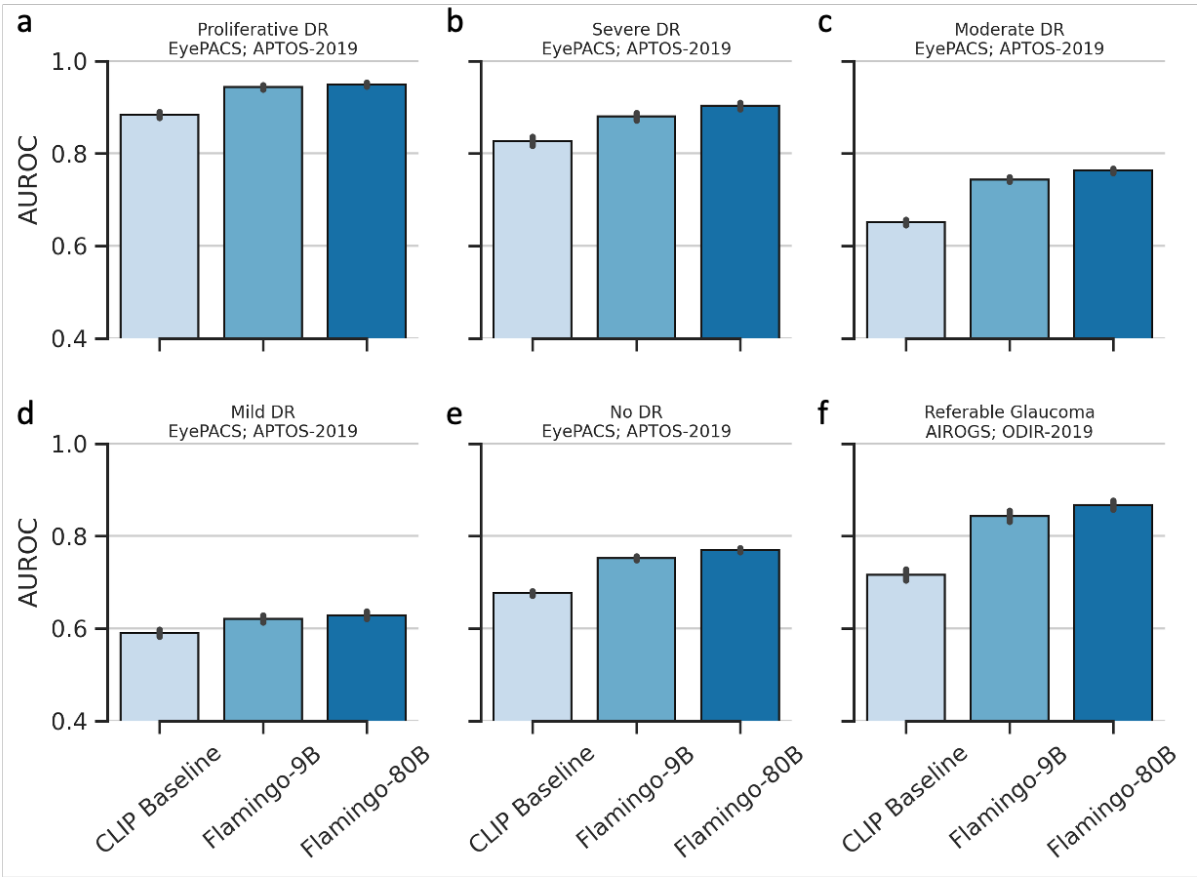


Figure 3: Performance in histopathological and dermatological image classification. (a-j) F1-score when classifying tissue type in task 1. Linear probes are fine-tuned on each dataset (Kather Colon and PanNuke) and evaluated on a hold-out test set. (a) to (i): classification of nine tissue types from colorectal cancer patients using image data from the Kather Colon dataset. (j): Malignancy classification in the PanNuke dataset in task 2. (k-r) AUC when classifying skin lesions. The probes are trained on the multimodal LLM's internal representations to predict the type of skin lesions (k-q) and malignancy (r). The center of each bar represents the mean of the metrics (F1 and AUC) and the error bars indicate the SDs. SDs and P-values are calculated using bootstrapping and paired, two-tailed t-tests.

Ophthalmology



Radiology

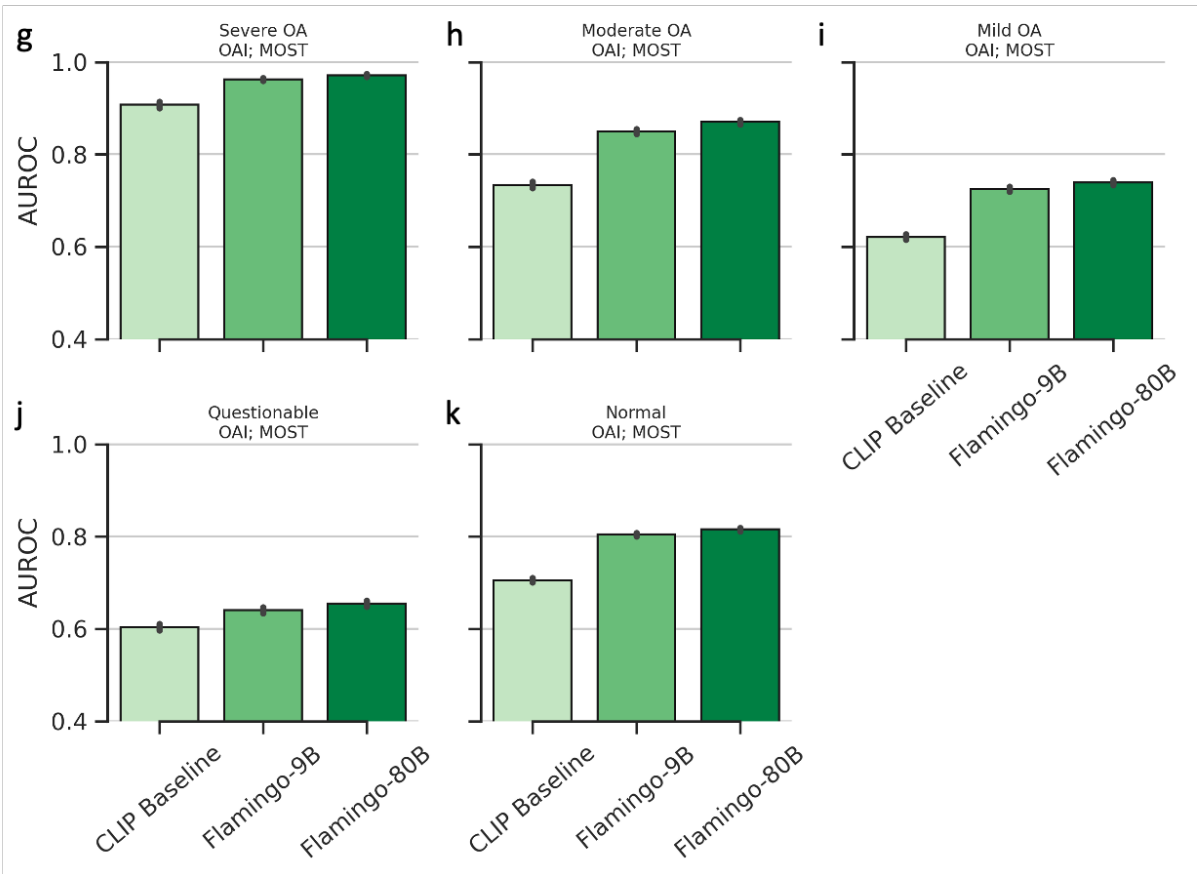


Figure 4: Performance in ophthalmological and radiological image classification. (a-e): Grading of diabetic retinopathy (DR). Linear probes are adapted to the EyePACS dataset by fine-tuning and evaluated on a hold-out test set to differentiate different stages of DR, such as proliferative DR, mild DR, and no DR eyes. (f): Classification of referrable glaucoma. (g-k) Performance in OA diagnosis based on knee radiographs. The center of each bar represents the mean AUC, and the error bars indicate the SDs. SDs and P-values are calculated using bootstrapping and paired, two-tailed t-tests.

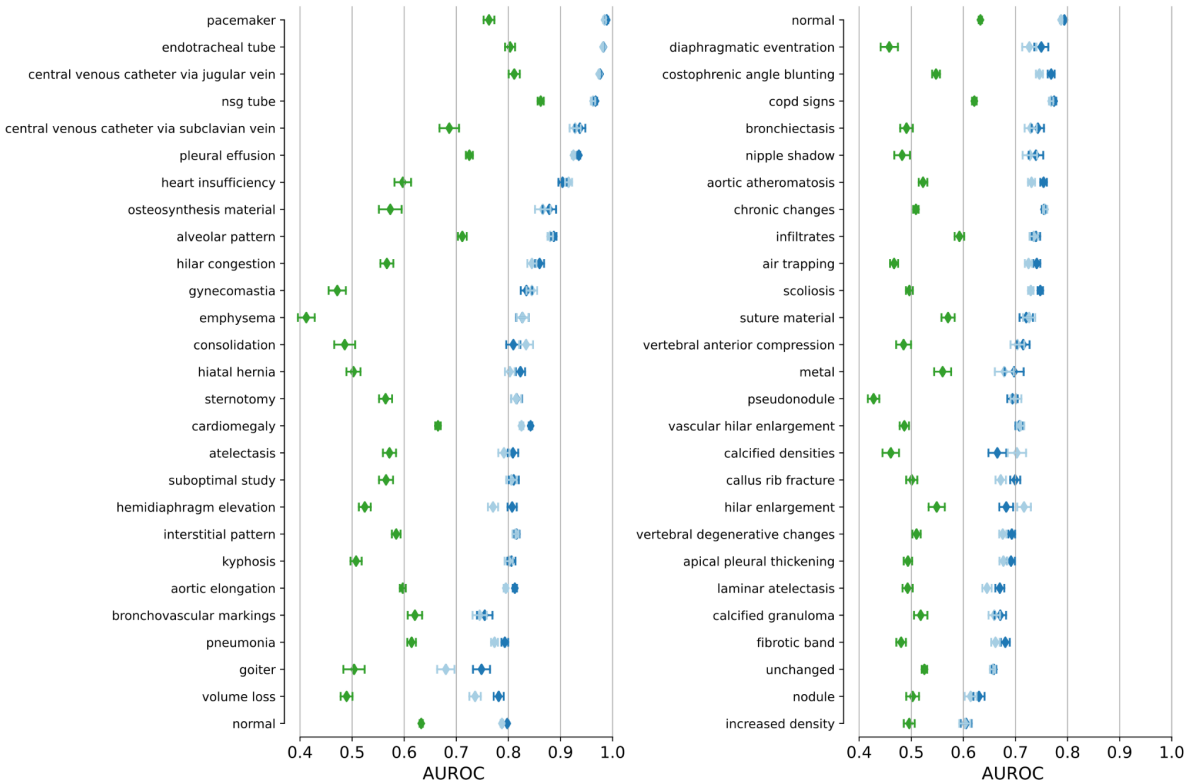


Figure 5: Detection of imaging and radiological findings on PadChest radiographs. Mean AUC and SD are shown for each finding with more than 50 entries in the PadChest testing cohort. The top 27 imaging findings are shown in the left panel and the remaining imaging findings are shown in the right panel. Flamingo-90B (green) consistently achieves higher AUC than CLIP (blue).

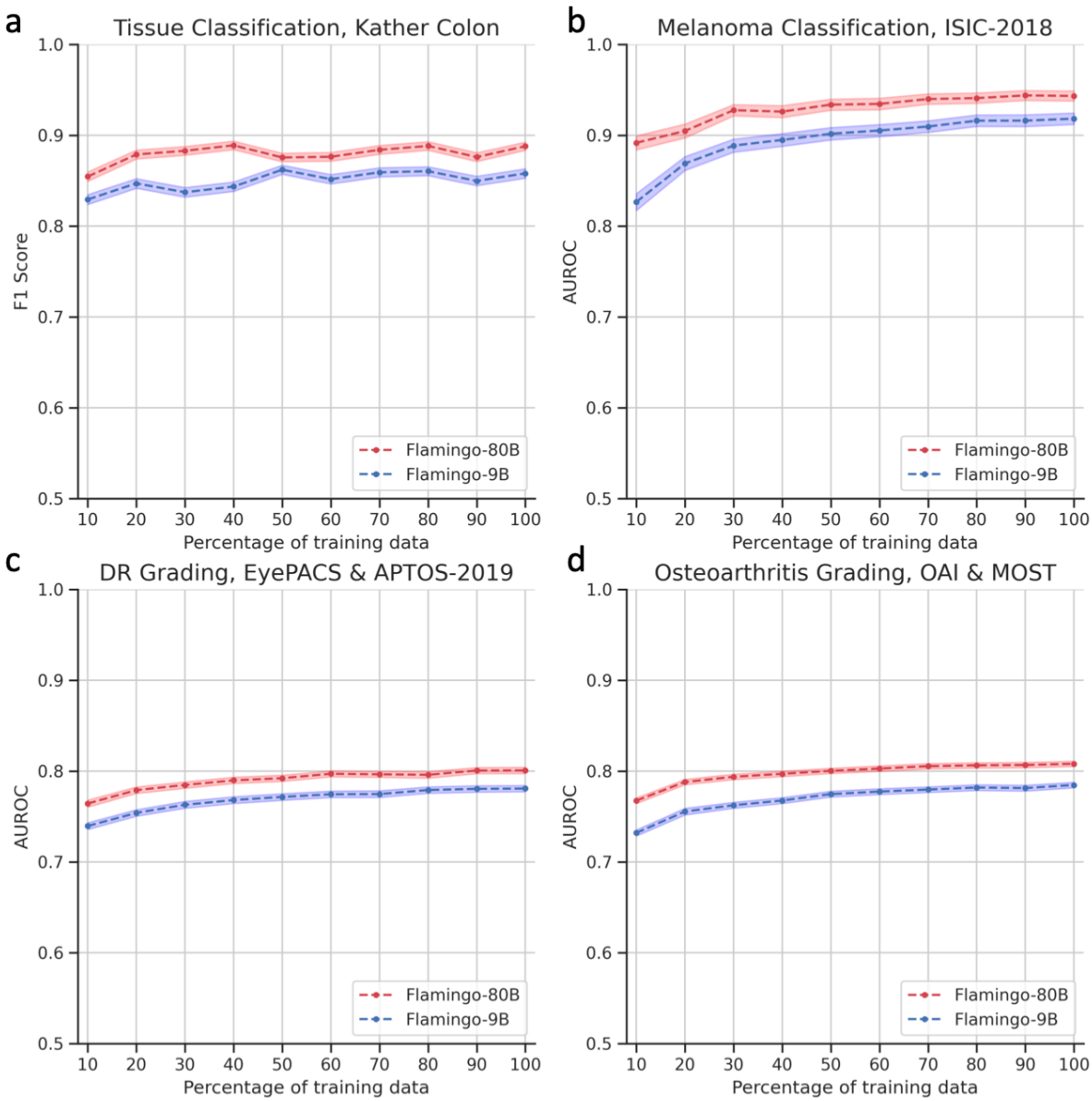


Figure 6: Robustness of our approach to data scarcity. In all four tasks, the performance of Flamingo-80B is robust to a reduced amount of training data. Tuning on only 10% of the training data, we maintained 95.8%, 94.3%, 95.2%, and 94.7% of the best performances in the pathology, dermatology, ophthalmology, and radiology tasks, respectively. The SDs of the AUC are plotted in colored bands, and the midpoints of the bands indicate the mean AUC.

Online Supplement

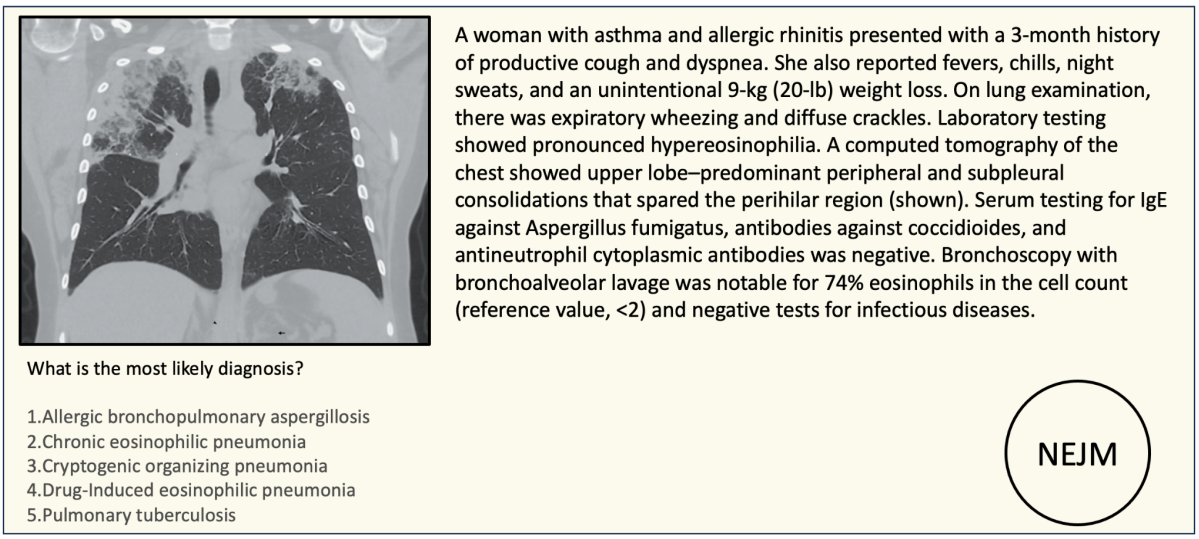


Figure S1: An illustrative example of the clinical case descriptions and answer choices from the “NEJM Image Challenge”.

Question: What is the diagnosis?
Choices:
1. Rocky mountain spotted fever
2. Hand foot and mouth disease
3. Infective endocarditis
4. Psoriasis
5. Secondary syphilis
Answer: Secondary syphilis
Reason: This symmetric, dusky red, and polymorphic papulosquamous rash is consistent with secondary syphilis.

Question: What process is illustrated in the radiograph?
Choices:
1. Paget's disease
2. Osteopetrosis
3. Hyperparathyroidism
4. Bone marrow hyperplasia
5. Acromegaly
Answer: Bone marrow hyperplasia
Reason: The "hair-on-end" appearance of the skull in the radiograph is consistent with massive bone marrow hyperplasia.

Question: This rash appeared following treatment for leukemia. What is the diagnosis?
Choices:
1. Cryoglobulinemia
2. Leukemia cutis
3. Herpes zoster
4. Graft-versus-host disease
5. Urticaria pigmentosa
Answer:

Figure S2: Two-shot example prompt used to query multimodal LLMs to answer NEJM Image Challenge questions.

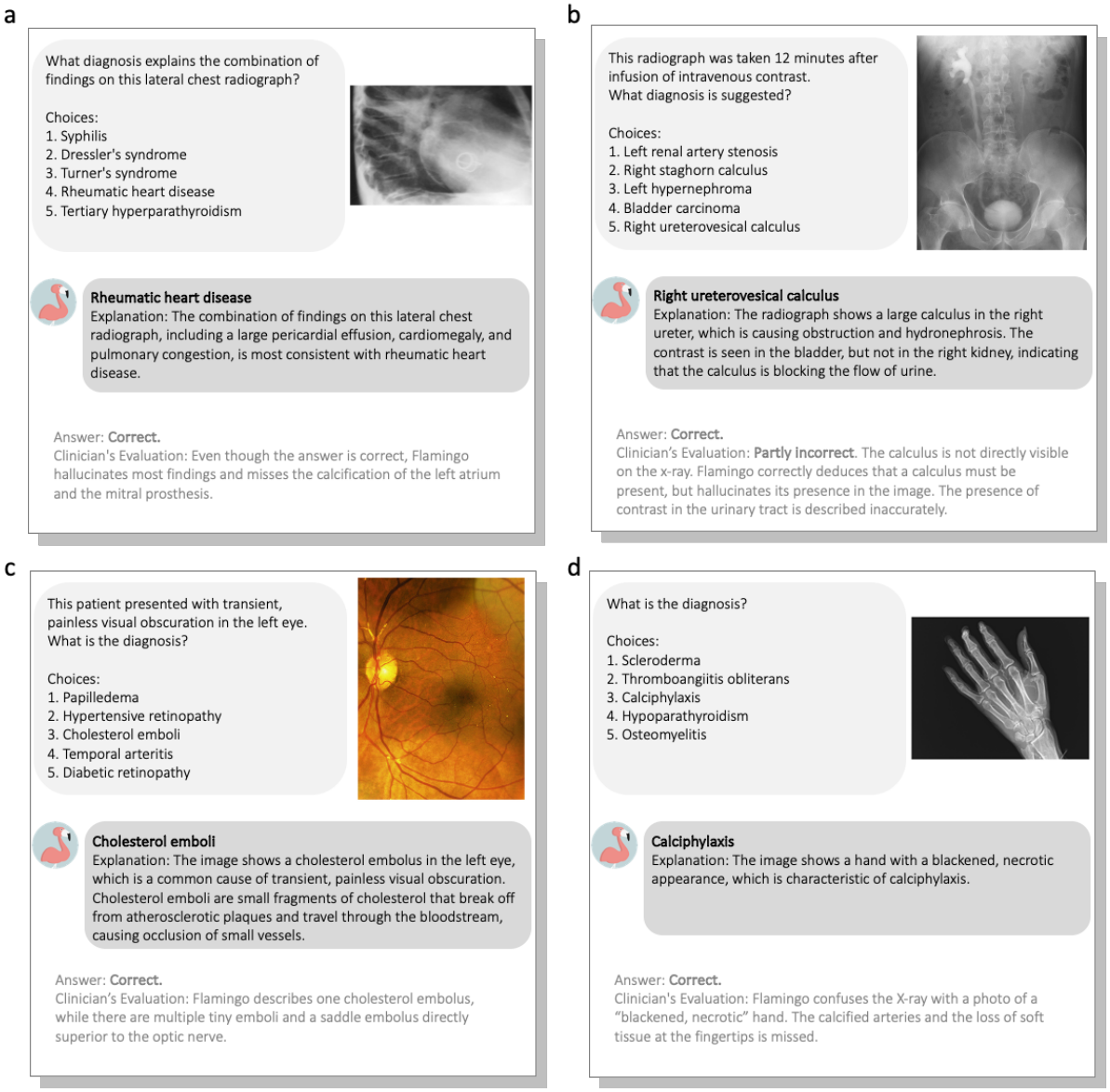


Figure S3: Cases from the NEJM Image Challenge with hallucinations. Flamingo-80B answered these questions correctly but reasoned incorrectly. We observed that multimodal LLMs can hallucinate strongly in certain medical cases such as (a), (c), (d), and (f).

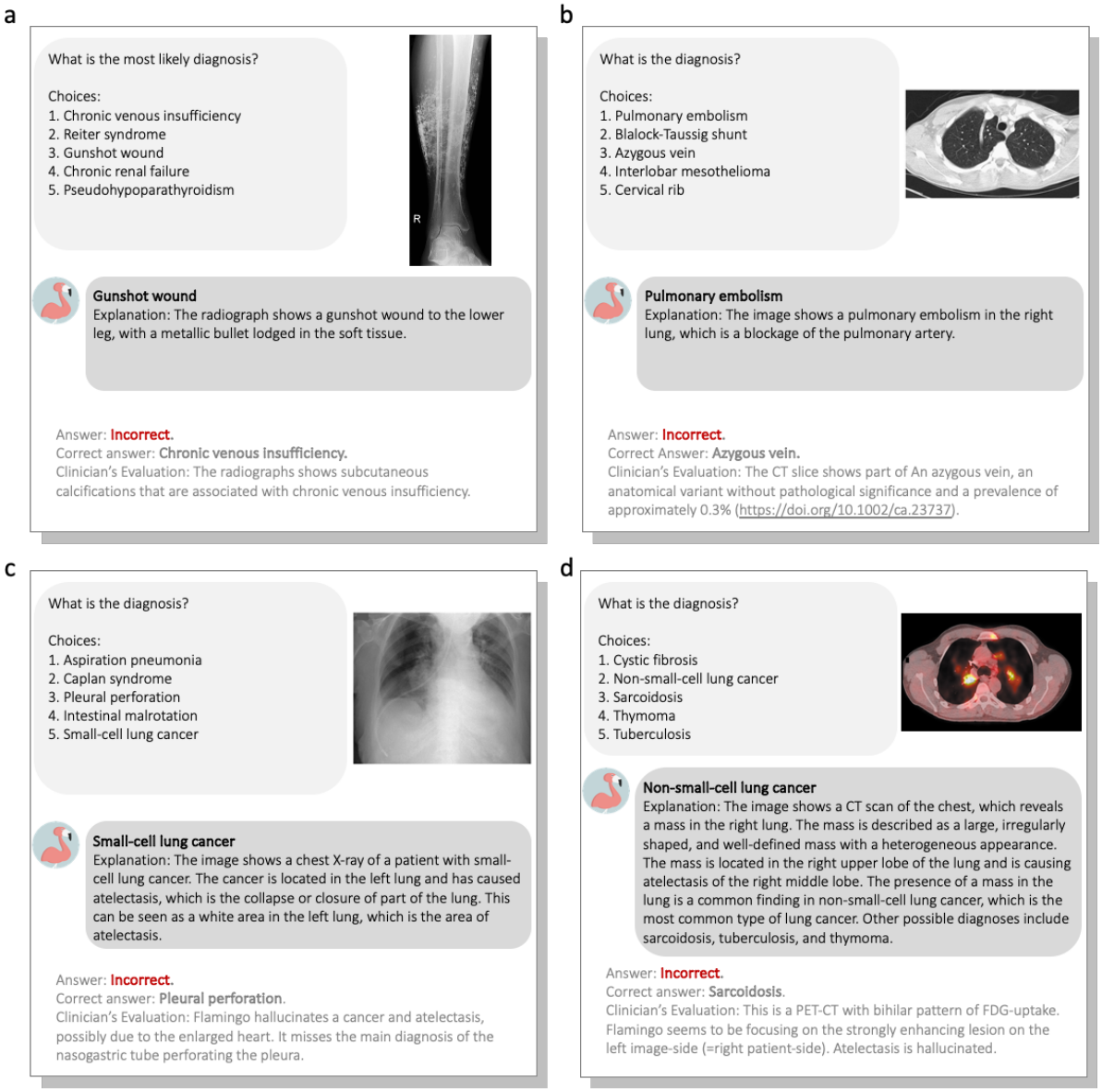


Figure S4: Selection of NEJM Image Challenge cases that were answered incorrectly. Flamingo-80B struggled to give the correct answer in these cases. We observe that Flamingo-80B mainly suffered from hallucinations (c), (e), and (f) or misperceptions (a), (b), and (d).

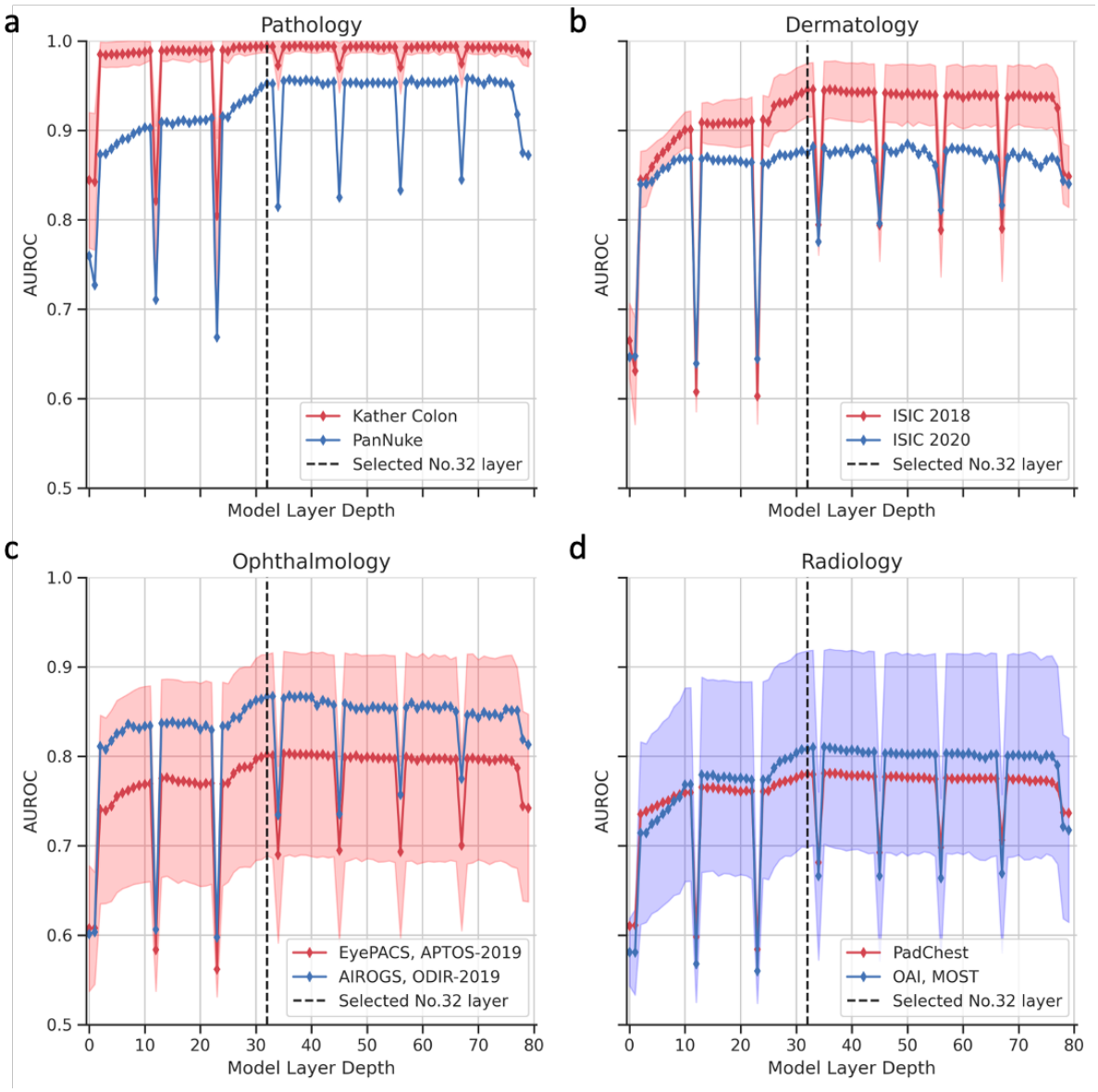


Figure S5: Testing the AUC for linear probes trained on each layer of the Flamingo-80B model. We select one layer (i.e., 32, highlighted in black, dashed lines) in a pre-experiment and then use it consistently for all subsequent experiments. In contrast to previously reported results,45 representations from the 80B multimodal LLM regularly fluctuate in quality across layers. We found that this phenomenon generalizes across evaluations in pathology (a), dermatology (b), ophthalmology (c), and radiology (d). The SDs of the AUCs are plotted in color bands, and the midpoints of the bands indicate the mean value of the AUC.

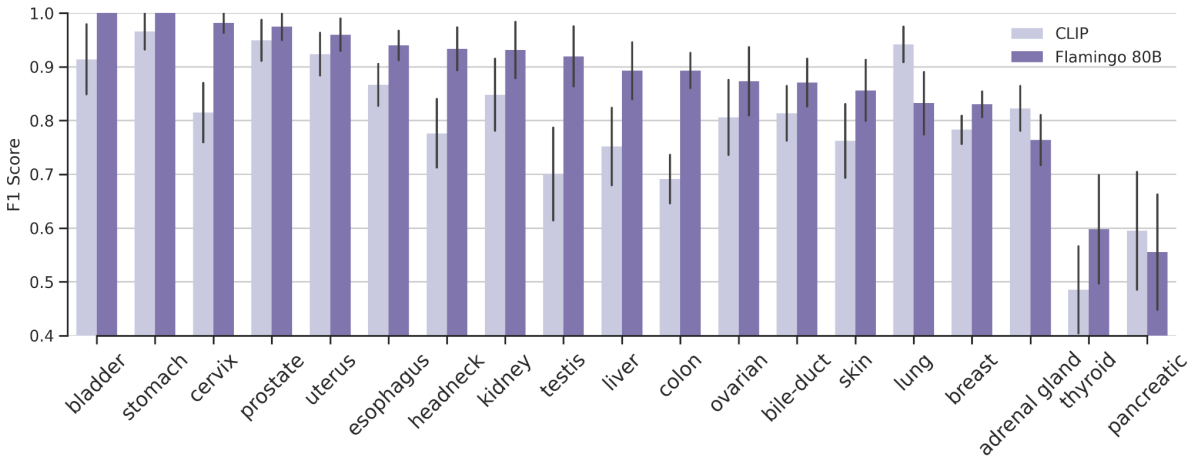


Figure S6: Evaluation of activation probes in the PanNuke dataset within each organ type.