Med-Flamingo: a Multimodal Medical Few-shot Learner

Michael Moor* CORRESPONDENCE TO: MDMOOR@CS.STANFORD.EDU
Qian Huang*
Shirley Wu
Michihiro Yasunaga
Yash Dalmia
Jure Leskovec†
Department of Computer Science, Stanford University, Stanford, USA

Cyril Zakka
Department of Cardiothoracic Surgery, Stanford Medicine, Stanford, USA

Eduardo Pontes Reis
Hospital Israelita Albert Einstein, São Paulo, Brazil

Pranav Rajpurkar
Department of Biomedical Informatics, Harvard Medical School, Boston, USA

Abstract

Medicine, by its nature, is a multifaceted domain that requires the synthesis of information across various modalities. Medical generative vision-language models (VLMs) make a first step in this direction and promise many exciting clinical applications. However, existing models typically have to be fine-tuned on sizeable down-stream datasets, which poses a significant limitation as in many medical applications data is scarce, necessitating models that are capable of learning from few examples in real-time. Here we propose Med-Flamingo, a multimodal few-shot learner adapted to the medical domain. Based on OpenFlamingo-9B, we continue pre-training on paired and interleaved medical image-text data from publications and textbooks. Med-Flamingo unlocks few-shot generative medical visual question answering (VQA) abilities, which we evaluate on several datasets including a novel challenging open-ended VQA dataset of visual USMLE-style problems. Furthermore, we conduct the first human evaluation for generative medical VQA where physicians review the problems and blinded generations in an interactive app. Med-Flamingo improves performance in generative medical VQA by up to 20% in clinician's rating and firstly enables multimodal medical few-shot adaptations, such as rationale generation. We release our model, code, and evaluation app.

Keywords: Medical AI; Few-shot learning; Foundation model; Vision-language model

1. Introduction

Large, pre-trained models (or foundation models) have demonstrated remarkable capabilities in solving an abundance of tasks by being provided only a few labeled examples as context (Bommasani et al., 2021). This is known as in-context learning (Brown et al., 2020), through which a model learns a task from a few provided examples specifically during prompting and without tuning the model parameters. In the medical domain, this bears great potential to vastly expand the capabilities of existing medical AI models (Moor et al., 2023). Most notably, it will enable medical AI models to handle the various rare cases faced by clinicians every day in a unified way, to provide relevant rationales to justify their statements, and to easily customize model generations to specific use cases.

Implementing the in-context learning capability in a medical setting is challenging due to the inherent complexity and multimodality of medical data and the diversity of tasks to be solved.

Previous efforts to create multimodal medical foundation models, such as ChexZero (Tiu et al., 2022) and BiomedCLIP (Zhang et al., 2023a), have made significant strides in their respective domains. ChexZero specializes in chest X-ray interpretation, while BiomedCLIP has been trained on more diverse images paired with captions from the biomedical literature.

* These authors contributed equally to this work.
† Last author.

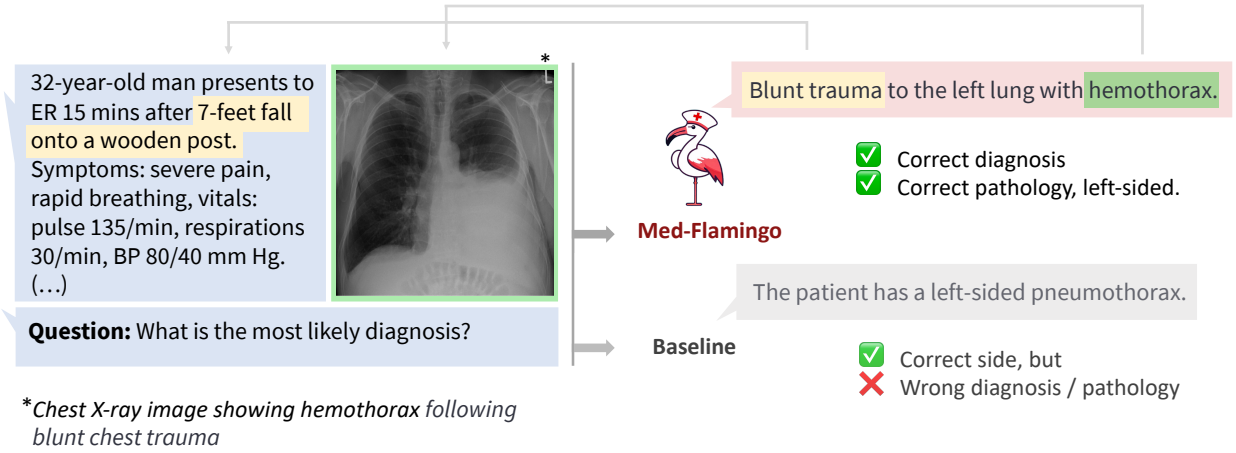

Figure 1: Example of how Med-Flamingo answers complex multimodal medical questions by generating open-ended responses conditioned on textual and visual information. The baseline response was given by the OpenFlamingo model, both models were few-shot prompted with 4 shots.

Other models have also been developed for electronic health record (EHR) data (Steinberg et al., 2021) and surgical videos (Kiyasseh et al., 2023). However, none of these models have embraced in-context learning for the multimodal medical domain. Existing medical VLMs, such as MedVINT (Zhang et al., 2023b), are typically trained on paired image-text data with a single image in the context, as opposed to more general streams of text that are interleaved with multiple images. Therefore, these models were not designed and tested to perform multimodal in-context learning with few-shot examples[1]

Here, we propose Med-Flamingo, the first medical foundation model that can perform multimodal in-context learning specialized for the medical domain. Med-Flamingo is a vision-language model based on Flamingo (Alayrac et al., 2022) that can naturally ingest data with interleaved modalities (images and text), to generate text conditioned on this multimodal input. Building on the success of Flamingo, which was among the first vision-language models to exhibit in-context learning and few-shot learning abilities, Med-Flamingo extends these capabilities to the medical domain by pre-training on multimodal knowledge sources across medical disciplines.

In preparation for the training of Med-Flamingo, our initial step involved constructing a unique, interleaved image-text dataset, which was derived from an extensive collection of over $4K$ medical textbooks (Section 3). Given the critical nature of accuracy and precision within the medical field, it is important to note that the quality, reliability, and source of the training data can considerably shape the results. Therefore, to ensure accuracy in medical facts, we meticulously curated our dataset from respected and authoritative sources of medical knowledge, as opposed to relying on potentially unreliable web-sourced data.

In our experiments, we evaluate Med-Flamingo on generative medical visual question-answering (VQA) tasks by directly generating open-ended answers, as opposed to scoring artificial answer options *ex post*—as CLIP-based medical vision-language models do. We design a new realistic evaluation protocol to measure the model generations' clinical usefulness. For this, we conduct an in-depth human evaluation study with clinical experts which results in a human evaluation score that serves as our main metric. In addition, due to existing medical VQA datasets being narrowly focused on image interpretation among the specialties of radiology and pathology, we create Visual USMLE[2], a challenging generative VQA dataset of complex USMLE-style problems across specialties, which are augmented with images, case vignettes, and potentially with lab results.

Averaged across three generative medical VQA datasets, few-shot prompted Med-Flamingo achieves the best average rank in clinical evaluation score (rank of 1.67, best prior model has 2.33), indicating that the model generates answers that are most preferred

1. For example, a challenge with multimodal in-context learning for existing medical vision language models is the potential for image information to leak across examples, potentially misleading the model.

2. USMLE stands for "United States Medical Licensing Examination".

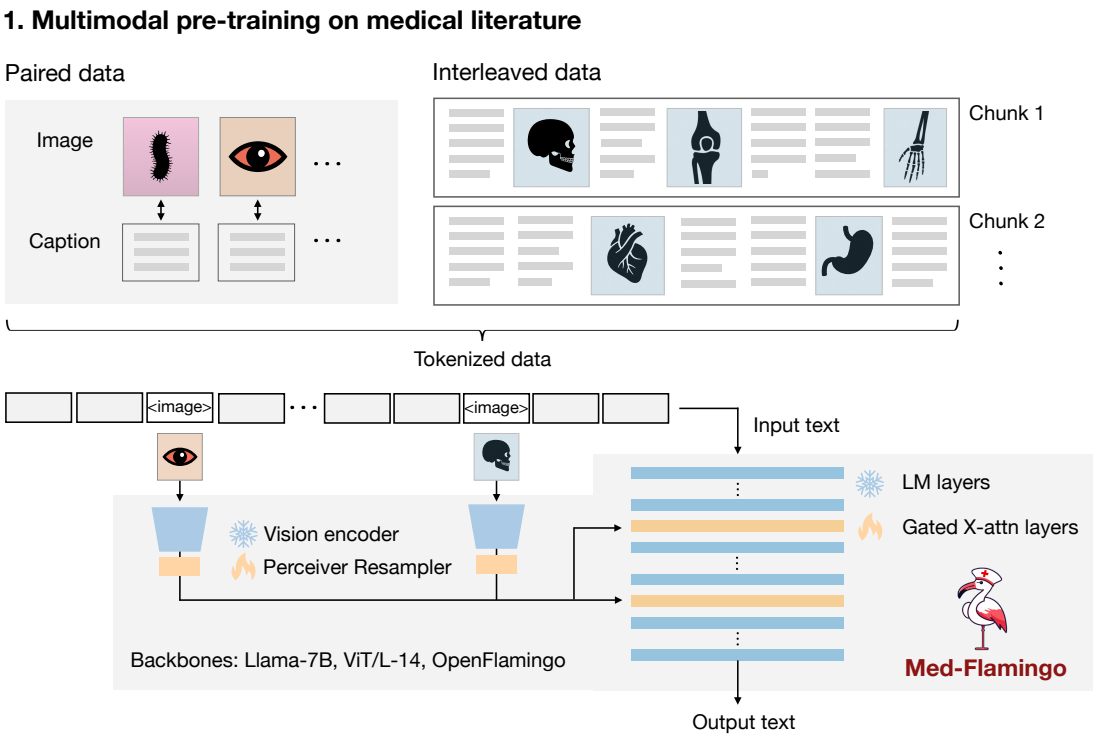

Figure 2: Overview of the Med-Flamingo model and the three steps of our study. First, we pre-train our Med-Flamingo model using paired and interleaved image-text data from the general medical domain (sourced from publications and textbooks). We initialize our model at the OpenFlamingo checkpoint continue pre-training on medical image-text data. Second, we perform few-shot generative visual question answering (VQA). For this, we leverage two existing medical VQA datasets, and a new one, Visual USMLE. Third, we conduct a human rater study with clinicians to rate generations in the context of a given image, question and correct answer. The human evaluation was conducted with a dedicated app and results in a clinical evaluation score that serves as our main metric for evaluation.

by clinicians, with up to 20% improvement over prior models. Furthermore, Med-Flamingo is capable of performing medical reasoning, such as answering complex medical questions (such as visually grounded USMLE-style questions) and providing explanations (i.e., rationales), a capability not previously demonstrated by other multimodal medical foundation models. However, it is important to note that Med-Flamingo's performance may be limited by the availability and diversity of training data, as well as the complexity of certain medical tasks. All investigated models and baselines would occasionally hallucinate or generate low-quality responses. Despite these limitations, our work represents a significant step forward in the development of multimodal

medical foundation models and their ability to perform multimodal in-context learning in the medical domain. We release the Med-Flamingo-9B checkpoint for further research, and make our code available under https://github.com/snap-stanford/med-flamingo. In summary, our paper makes the following contributions:

1. We present the first multimodal few-shot learner adapted to the medical domain, which promises novel clinical applications such as rationale generation and conditioning on retrieved multimodal context.
2. We create a novel dataset that enables the pre-training of a multimodal few-shot learner for the general medical domain.
3. We create a novel USMLE-style evaluation dataset that combines medical VQA with complex, across-specialty medical reasoning.
4. We highlight shortcomings of existing evaluation strategies, and conduct an in-depth clinical evaluation study of open-ended VQA generations with medical raters using a dedicated evaluation app.

2. Related works

The success of large language models (LLMs) (Brown et al.; Liang et al., 2022; Qin et al., 2023) has led to significant advancements in training specialized models for the medical domain. This has resulted in the emergence of various models, including BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), PubMedBERT (Gu et al., 2021), BioLinkBERT (Yasunaga et al., b), DRAGON (Yasunaga et al., a), BioMedLM (Bolton et al., 2022), BioGPT (Luo et al., 2022), and Med-PaLM (Singhal et al., 2022). Although these medical language models are typically smaller than general-purpose LLMs like GPT-3 (Brown et al.), they can match or even surpass their performance on medical tasks, such as medical question answering.

Recently, there has been a growing interest in extending language models to handle vision-language multimodal data and tasks (Su et al., 2019; Ramesh et al.; Alayrac et al., 2022; Aghajanyan et al.; Yasunaga et al., 2023). Furthermore, many medical applications involve multimodal information, such as radiology tasks that require the analysis of both X-ray images and radiology reports (Tiu et al., 2022). Motivated by these factors, we present a medical vision-language model (VLM). Existing medical VLMs include BiomedCLIP (Zhang et al., 2023a), MedVINT (Zhang et al., 2023b). While BiomedCLIP is an encoder-only model, our focus lies in developing a generative VLM, demonstrating superior performance compared to MedVINT. Finally, Llava-Med is another recent medical generative VLM (Li et al., 2023), however the model was not yet available for benchmarking.

3. Med-Flamingo

To train a Flamingo model adapted to the medical domain, we leverage the pre-trained OpenFlamingo-9B model checkpoint (Awadalla et al., 2023), which is a general-domain VLM that was built on top of the frozen language model LLaMA-7B (Touvron et al., 2023) and frozen vision encoder CLIP ViT/L-14 (Radford et al.). We perform continued pre-training in the medical domain which results in the model we refer to as Med-Flamingo.

3.1. Data

We pre-train Med-Flamingo by jointly training on interleaved image-text data and paired image-text data. As for the interleaved dataset, we created a interleaved dataset from a set of medical textbooks, which we subsequently refer to as MTB. As for the paired datasets, we used PMC-OA (Lin et al., 2023).

MTB We construct a new multimodal dataset from a set of 4721 textbooks from different medical specialties (see Figure 3). During preprocessing, each book is first converted from PDF to HTML with all tags removed, except the image tags are converted to <image> tokens. We then carry out data cleaning via deduplication and content filtering. Finally, each book with cleaned text and images is then chopped into segments for pretraining so that each segment contains at least one image and up to 10 images and a maximum length. In total, MTB consists of approximately 0.8M images and 584M tokens. We use 95% of the data for training and 5% of the data for evaluation during the pre-training.

PMC-OA We adopt the PMC-OA dataset (Lin et al., 2023) which is a biomedical dataset with 1.6M image-caption pairs collected from PubMedCentral's OpenAccess subset. We use 1.3M image-caption pairs for training and 0.16M pairs for evaluation following the public split[3].

3.2. Objectives

We follow the original Flamingo model approach (Alayrac et al.), which considers the following language modelling problem:

3. https://huggingface.co/datasets/axiong/pmc_oa_beta

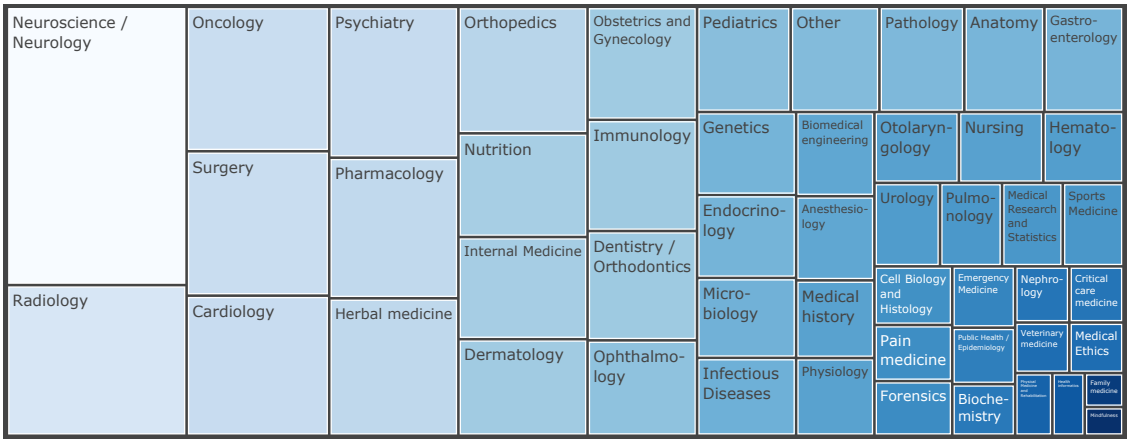

Figure 3: Overview of the distribution of medical textbook categories of the MTB dataset. We classify each book title into one of the 49 manually created categories or "other" using the Claude-1 model.

$$p(y_\ell \,|\, x_{<\ell}, y_{<\ell}) = \prod_{\ell=1}^{L} p(y_\ell \,|\, y_{<\ell}, x_{<\ell}),$$

where y_ℓ refers to the ℓ-th language token, $y_{<\ell}$ to the set of preceding language tokens, and $x_{<\ell}$ to the set of preceding visual tokens. As we focus on modelling the medical literature, here we consider only image-text data (i.e., no videos).

Following Alayrac et al., we minimize a joint objective $\mathcal{L}$ over paired and interleaved data:

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim D_p}[S(x,y)] + \lambda \cdot \mathbb{E}_{(x,y)\sim D_i}[S(x,y)],$$

where $S(x,y) = -\sum_{\ell=1}^{L} \log p\,(y_\ell \,|\, y_{<\ell}, x_{<\ell})$, and D_p and D_i stand for the paired and interleaved dataset, respectively. In our case, we use $\lambda = 1$.

3.3. Training

We performed multi-gpu training on a single node with 8x 80GB NVIDIA A100 GPUs. We trained the model using DeepSpeed ZeRO Stage 2: Optimizer states and gradients are sharded across devices. To further reduce memory load, we employed the 8-bit AdamW optimizer as well as the memory-efficient attention implementation of PyTorch 2.0. Med-Flamingo was initialized at the checkpoint of the Open-Flamingo model and then pre-trained for 2700 steps (or 6.75 days in wall time, including the validation steps), using 50 gradient accumulation steps and a per-device batch size of 1, resulting in a total batch size of 400. The model has $1.3B$ trainable parameters (gated cross attention layers and perceiver layers) and roughly $7B$ frozen parameters (decoder layers and vision encoder), which results in a total of $8.3B$ parameters. Note that this is the same number parameters as in the OpenFlamingo-9B model (version 1).

4. Evaluation

4.1. Automatic Evaluation

Baselines To compare generative VQA abilities against the literature, we consider different variants of the following baselines:

1. MedVINT (Zhang et al., 2023b), a visual instruction-tuned VLM based on Llama. As this model was not designed to do few-shot learning (e.g. the image information is prepended to the overall input), we report two modes for MedVINT: zero-shot and fine-tuned, where the model was fine-tuned on the training split of the VQA dataset. Since the rather small Visual-USMLE dataset has no separate training split, we ommit the fine-tuned baseline for that dataset. We used the MedVInT-TD model with PMC-LLaMA and PMC-CLIP backbones.

2. OpenFlamingo (Awadalla et al., 2023), a powerful VLM which was trained on general-domain data, and which served as the base model to train Med-Flamingo. We report both zero-shot and few-shot performance. We expect Flamingo-type models to shine in the few-shot setting which they are designed for (as already the pre-training task includes multiple interleaved image-text examples).

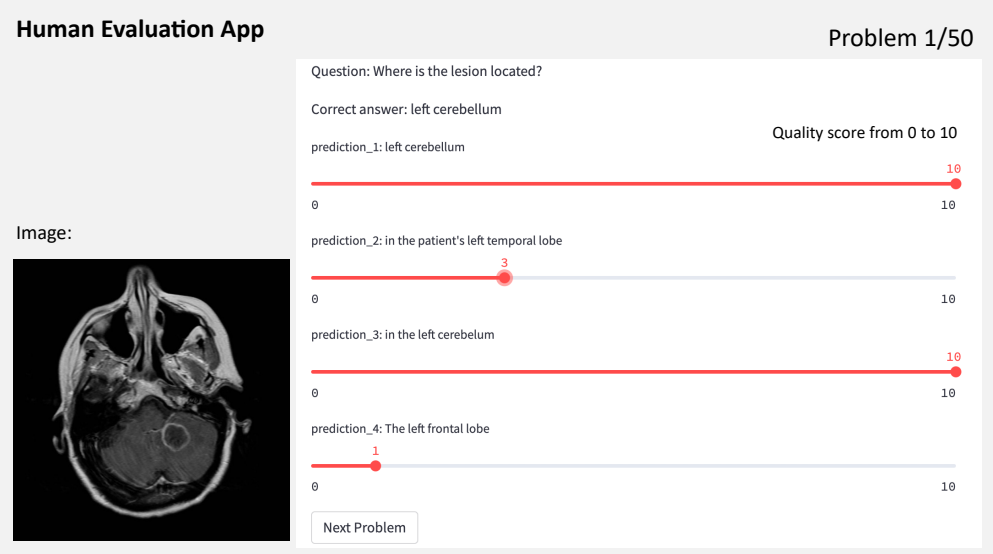

Figure 4: Illustration of our Human evaluation app that we created for clinical experts to evaluate generated answers.

Evaluation datasets To evaluate our model and compare it against the baselines, we leverage two existing VQA datasets from the medical domain (VQA-RAD and PathVQA). Upon closer inspection of the VQA-RAD dataset, we identified severe data leakage in the official train / test splits, which is problematic given that many recent VLMs fine-tune on the train split. To address this, we created a custom train / test split by seperately splitting images and questions (each 90% / 10%) to ensure that no image or question of the train split leaks into the test split. On these datasets, 6 shots were used for few-shot, whereas as the in-context examples were randomly drawn from the respective train splits.

Furthermore, we create Visual USMLE, a challenging multimodal problem set of 618 USMLE-style questions which are not only augmented with images but also with a case vignette and potentially tables of laboratory measurements. The Visual USMLE dataset was created by adapting problems from the Amboss platform (using licenced user access). To make the Visual USMLE problems more actionable and useful, we rephrased the problems to be open-ended instead of multiple-choice. This makes the benchmark harder and more realistic, as the models have to come up with differential diagnoses and potential procedures completely on their own—as opposed to selecting the most reasonable answer choice from few choices. Figure 7 gives an overview of the broad range of specialties that are covered in the dataset, greatly extending existing medical VQA datasets which are narrowly focused on radiology and pathology. For this comparatively small dataset, instead of creating a training split for finetuning, we created a small train split of 10 problems which can be used for few-shot prompting. For this dataset (with considerably longer problems and answers), we used only 4 shots to fit in the context window.

Evaluation metrics Previous works in medical vision-language modelling typically focused scoring all available answers of a VQA dataset to arrive at a classification accuracy. However, since we are interested in *generative* VQA (as opposed to post-hoc scoring different potential answers), for sake of clinical utility, we employ the following evaluation metrics that directly assess the quality of the generated answer:

1. Clinical evaluation score, as rated by three medical doctors (including one board-certified radiologist) using a human evaluation app that we developed for this study. More details are provided in Section 4.2.

2. BERT similarity score (BERT-sim), the F1 BERT score between the generated answer and the correct answer (Zhang et al., 2020).

3. Exact-match, the fraction of generated answers that exactly match (modulo punctuation) the correct answer. This metric is rather noisy and conservative as useful answers may not lexically match the correct answer.

Dataset	Model	Clinical eval. score	BERT-sim	Exact-match
VQA-RAD	MedVINT zero-shot	4.63	0.628	0.167
	MedVINT fine-tuned ($\sim 2K$ samples)	2.87	0.611	0.133
	OpenFlamingo zero-shot	4.39	0.490	0.000
	OpenFlamingo few-shot	4.69	0.645	**0.200**
	Med-Flamingo zero-shot	3.82	0.480	0.000
	Med-Flamingo few-shot	**5.61**	**0.650**	**0.200**
Path-VQA	MedVINT zero-shot	0.13	0.608	0.272
	MedVINT fine-tuned ($\sim 20K$ samples)	1.23	**0.723**	**0.385**
	OpenFlamingo zero-shot	**2.16**	0.474	0.009
	OpenFlamingo few-shot	2.08	0.669	0.288
	Med-Flamingo zero-shot	1.72	0.521	0.120
	Med-Flamingo few-shot	1.81	0.678	0.303
Visual USMLE	MedVINT zero-shot	0.41	0.421	-
	OpenFlamingo zero-shot	4.31	**0.512**	-
	OpenFlamingo few-shot	3.39	0.470	-
	Med-Flamingo zero-shot	4.18	0.473	-
	Med-Flamingo few-shot	**4.33**	0.431	-

Table 1: Performance metrics across VQA-Rad, PathVQA, and Visual USMLE datasets. Best scores are highlighted in bold. Emphasis is placed on the clinical evaluation score. BERT-sim likely does not capture all fine-grained medical details. Exact-match is brittle, though provides a conservative measure. Exact-match was uninformative (constant 0) for Visual USMLE due to long correct answers. The fine-tuned baseline did not surpass zero-shot performance in VQA-Rad, possibly due to its small size and custom splits to prevent leakage. Notably, the PathVQA dataset revealed a pronounced performance deficit in pathology, underscoring that prior classification metrics might have overestimated VLMs' efficacy in this domain.

4.2. Human evaluation

We implemented a human evaluation app using Streamlit to visually display the generative VQA problems for clinical experts to rate the quality of the generated answers with scores from 0 to 10. Figure 4 shows an examplary view of the app. For each VQA problem, human raters are provided with the image, the question, the correct answer, and a set of blinded generations (e.g., appearing as "prediction_1" in Figure 4), that appear in randomized order. As for human raters, we employed three medical doctors that were affiliated with the same academic center, however received their medical training in different countries. The team of raters included one board-certified radiologist.

4.3. Deduplication and leakage

During the evaluation of the Med-Flamingo model, we were concerned that there may be leakage between the pre-training datasets (PMC-OA and MTB) and the down-stream VQA datasets used for evaluation; this could inflate judgements of model quality, as the model could memorize image-question-answer triples.

To alleviate this concern, we performed data deduplication based upon pairwise similarity between images from our pre-training datasets and the images from our evaluation benchmarks. To detect similar images, in spite of perturbations due to cropping, color shifts, size, etc, we embedded the images using Google's Vision Transformer, preserving the last hidden state as the resultant embedding (Dosovitskiy et al., 2021). We then found the k-nearest neighbors to each evaluation image from amongst the pre-training images (using the FAISS library) (Johnson et al., 2019). We then sorted and visualized image-image pairs by least euclidean distance; we found that images might be duplicates until a pairwise distance value of 80; beyond this point, there were no duplicates.

This process revealed that the pretraining datasets leaked into the PVQA evaluation benchmark. Out of 6700 total images in PVQA test set, we judged 194 to be highly similar to images in the pretraining datasets, and thus, we removed them from our down-stream evaluation.

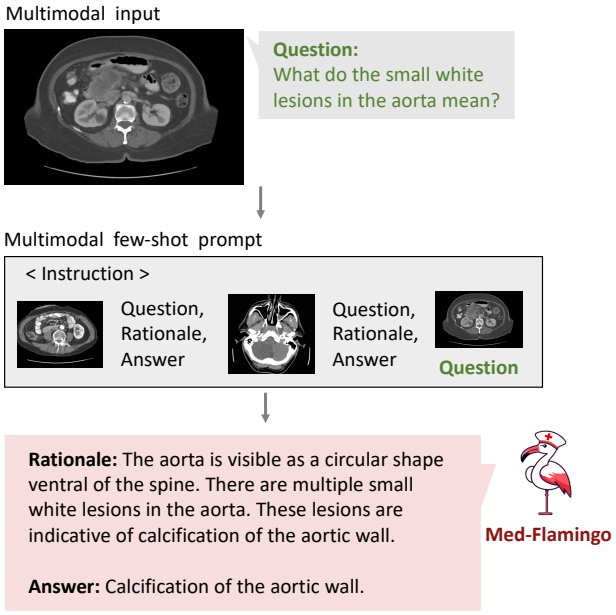

Figure 5: Multimodal medical few-shot prompting illustrated with an example. Few-shot prompting here allows users to customize the response format, *e.g.*, to provide rationales for the provided answers. In addition, multimodal few-shot prompts potentially offer the ability to include relevant context retrieved from the medical literature.

5. Results

In our experiments, we focus on generative medical visual question answering (VQA). While recent medical VLMs predominantly performed VQA in a non-generative but rather discriminative manner (i.e., by scoring different answer choices), we believe that this ex-post classification to carry less clinical usefulness, than directly generating responses. On the other hand, generative VQA is more challenging to evaluate, as automated metrics suffer from significant limitations as they do not fully capture the domain-specific context. Thus, we perform a human evaluation study where clinical experts review model generations (blinded) and score them (between 0 and 10) in terms of clinical usefulness.

Conventional VQA datasets Table 1 shows the results for VQA-RAD, the radiological VQA dataset for which we created custom splits to address leakage (see Section4). Med-Flamingo few-shot shows strong results, improving the clinical eval score by $\sim 20\%$ over the best baseline. In this dataset, the auxiliary metrics are rather aligned with clinical preference. Finetuning the MedVINT baseline did not lead to improved performance on this dataset which may be due to its small size. MedVINT zero-shot outperforms the other zero-shot ablations which may be partially attributed to its instruction tuning step on PMC-VQA.

Table 1 shows for the results for Path-VQA, the pathology VQA dataset. Compared to the other datasets, all models overall perform poorer on the Path-VQA dataset in terms of clinical evaluation score. We hypothesize that this has to do with the fact the models are not pre-trained on actual large-scale and fine-grained pathology image datasets, but only on a rather small amount of pathology literature (which may not be enough to achieve strong performance). For instance, Figure 3 shows that only a small fraction of our training data covers pathology. In the automated metrics (BERT-sim and exact-match), Med-Flamingo improves upon the OpenFlamingo baseline, however the overall quality does not improve (as seen in the clinical evaluation score). MedVINT was fine-tuned on a sizeable training split which results in strong automated metrics, but did not result in a clinical evaluation score that matches any Flamingo variant.

Visual USMLE Table 1 shows the results for the Visual USMLE dataset. Med-Flamingo (few-shot) results in the clinically most preferrable generations, whereas OpenFlamingo (zero-shot) is a close runner-up. As the ground truth answers were rather lengthy paragraphs, exact match was not an informative metric (constant 0 for all methods). The few-shot prompted models lead to lower automated scores than their zero-shot counterparts, which we hypothesize has to do with the fact that the USMLE problems are long (long vignettes as well as long answers) which forced us to summarize the questions and answers when designing few-shot prompts (for which we used GPT-4). Hence, it's possible that those prompts lead to short answers that in terms of BERT-sim score may differ more from the correct answer than a more wordy zero-shot generation.

Across datasets Overall, we find that Med-Flamingo's multimodal in-domain few-shot learning abilities lead to favorable generative VQA performance, leading to the lowest average rank of 1.67 in terms of clinical evaluation score as averaged across all evaluation datasets. As runner-up, OpenFlamingo zero-shot achieves a rank of 2.33.

Qualitative analysis Finally, we showcase few examples of Med-Flamingo generations in more detail in Figures 1,5, and 8. Figure 5 exemplifies that a medical few-shot learner like Med-Flamingo can be prompted to

generate rationale for its VQA answer. The shown example is impressive in that the rationale is visually guiding the reader towards the object of interest (calcification of the aortic wall). We note, however, that at this stage, few-shot multimodal prompted rationales may not be robust, especially when a model arrives at a wrong answer.

Figures 1 and 8 showcase two example problems from the Visual USMLE dataset. The problem descriptions were slightly rephrased and summarized using GPT-4 for display. In Figure 8, Med-Flamingo generates the correct answer while not mentioning the underlying diagnosis (urothelial cancer) as it was not asked for. By contrast, we observed baselines to directly diagnose the patient (instead of answering the actual question in a targeted way). The problem in Figure 1 illustrates that Med-Flamingo has the ability to integrate complex medical history information together with visual information to synthesize a comprehensive diagnosis that draws from the information of both modalities. As for failure modes, we occasionally observed that information from the in-context examples can leak into the final generation.

6. Discussion

In this paper, we presented Med-Flamingo, the first medically adapted multimodal few-shot learner. While this is an early proof-of-concept for a medical multimodal few-shot learner, we expect to see significant improvements with increased model and data scale, more thoroughly cleaned data, as well as with alignment to human preference via instruction tuning or explicit optimization for preferences.

We expect that the rise of multimodal medical few-shot learners will lead to exciting opportunities with regard to model explainability (via rationale generation) as well as grounding the model in verified sources (via multimodal retrieval to augment the few-shot prompt). Thereby, our work serves as a first step towards more generalist medical AI models (Moor et al., 2023).

Limitations This work demonstrates a proof-of-concept. As such, Med-Flamingo is *not* intended nor safe for clinical use. In all VLMs we analyzed, hallucinations were observed. Furthermore, as Med-Flamingo is a pre-trained model without further instruction or preference tuning, it is possible that the model occasionally outputs low-quality generations.

Future work It will be an exciting route for future work to further train Med-Flamingo on clinical data, high-resolution medical image datasets as well as 3D volumes and medical videos. While current general-purpose medical VLMs are pre-trained on the broad medical literature (*i.e.,* they are only "book-smart"), also learning from diverse patient data directly will become crucial for down-stream applications.

Acknowledgments

We thank Rok Sosič for his technical support in the data preprocessing.

References

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. URL http://arxiv.org/abs/2201.07520.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*. URL https://openreview.net/forum?id=EbMuimAbPbs.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL https://doi.org/10.5281/zenodo.7733589.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. BioMedLM: a domain-specific large language model for biomedical

text. 2022. URL https://crfm.stanford.edu/2022/12/15/biomedlm.html.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Dani Kiyasseh, Runzhuo Ma, Taseen F Haque, Brian J Miles, Christian Wagner, Daniel A Donoho, Animashree Anandkumar, and Andrew J Hung. A vision transformer for decoding surgeon activity from surgical videos. *Nature Biomedical Engineering*, pages 1–17, 2023.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. URL https://proceedings.mlr.press/v139/radford21a.html. ISSN: 2640-3498.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. URL https://proceedings.mlr.press/v139/ramesh21a.html. ISSN: 2640-3498.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL http://arxiv.org/abs/2212.13138.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, a. URL https://openreview.net/forum?id=4NpoSrT8uU-.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016. Association for Computational Linguistics, b. doi: 10.18653/v1/2022.acl-long.551. URL https://aclanthology.org/2022.acl-long.551.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023a.

Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

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

Appendix A. Appendix

A.1. Details for MTB dataset

Clustering the images In a post-hoc analysis, we clustered the image embeddings of the MTB dataset into a large number of clusters (100) and manually reviewed examples of each cluster to assign an annotation. We discard noisy or unclear clusters and display the remaining clusters and their frequency in Figure 6.

Classification of book titles Here, we provide further details about the creation of Figure 3. Table 2 lists the categories used to prompt the Claude-1 model to classify each book title. We initially prompted with 3 more very rare categories (Geriatrics, Occupational medicine, Space medicine), but merge them into the "Other" group for visualization purposes.

A.2. Details for Visual USMLE dataset

Figure 7 shows the distribution of specialty topics among the problems of the Visual USMLE dataset. Again, we used Claude-1 to classify each problem into categories provided in Table 2.

In Figure 8, we display an example problem of the Visual USMLE dataset.

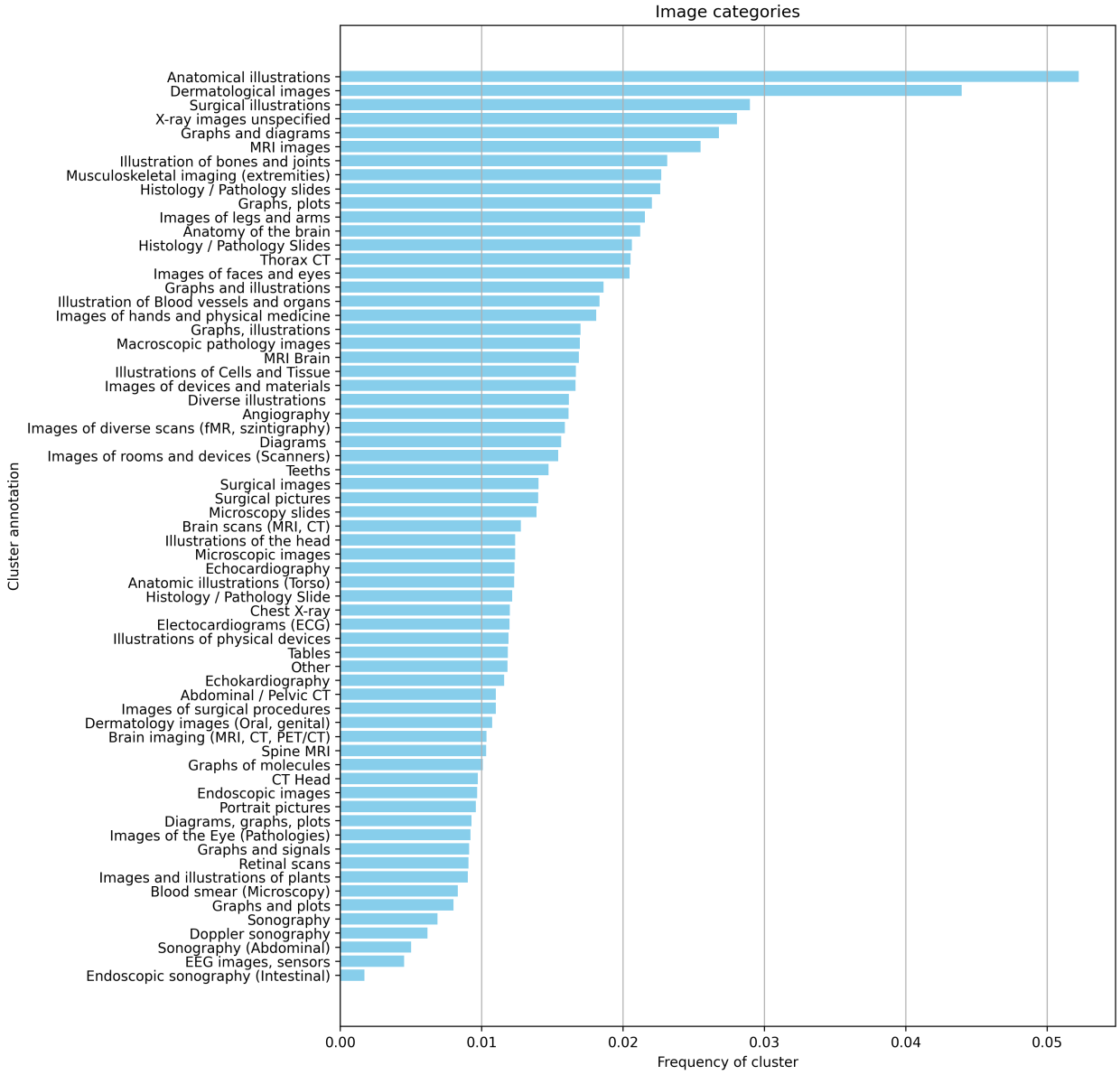

Figure 6: Distribution of manually annotated image clusters in the MTB dataset.

Neuroscience/Neurology	Obstetrics and Gynecology	Infectious Diseases
Radiology	Dermatology	Family medicine
Oncology	Immunology	Biomedical engineering
Surgery	Dentistry / Orthodontics	Anesthesiology
Cardiology	Ophthalmology	Physiology
Psychiatry	Pediatrics	Medical history
Pharmacology	Pathology	Nursing
Herbal medicine	Anatomy	Otolaryngology
Orthopedics	Gastroenterology	Hematology
Nutrition	Endocrinology	Urology
Internal Medicine	Genetics	Pulmonology
Sports Medicine	Medical Research and Statistics	Emergency Medicine
Cell Biology and Histology	Pain medicine	Public Health and Epidemiology
Forensics	Biochemistry	Nephrology
Critical care medicine	Medical Ethics	Veterinary medicine
Physical Medicine and Rehabilitation	Health informatics	Mindfulness
Other		

Table 2: List of 49 Categories (and "Other") used for visualing the MTB dataset in Figure 3

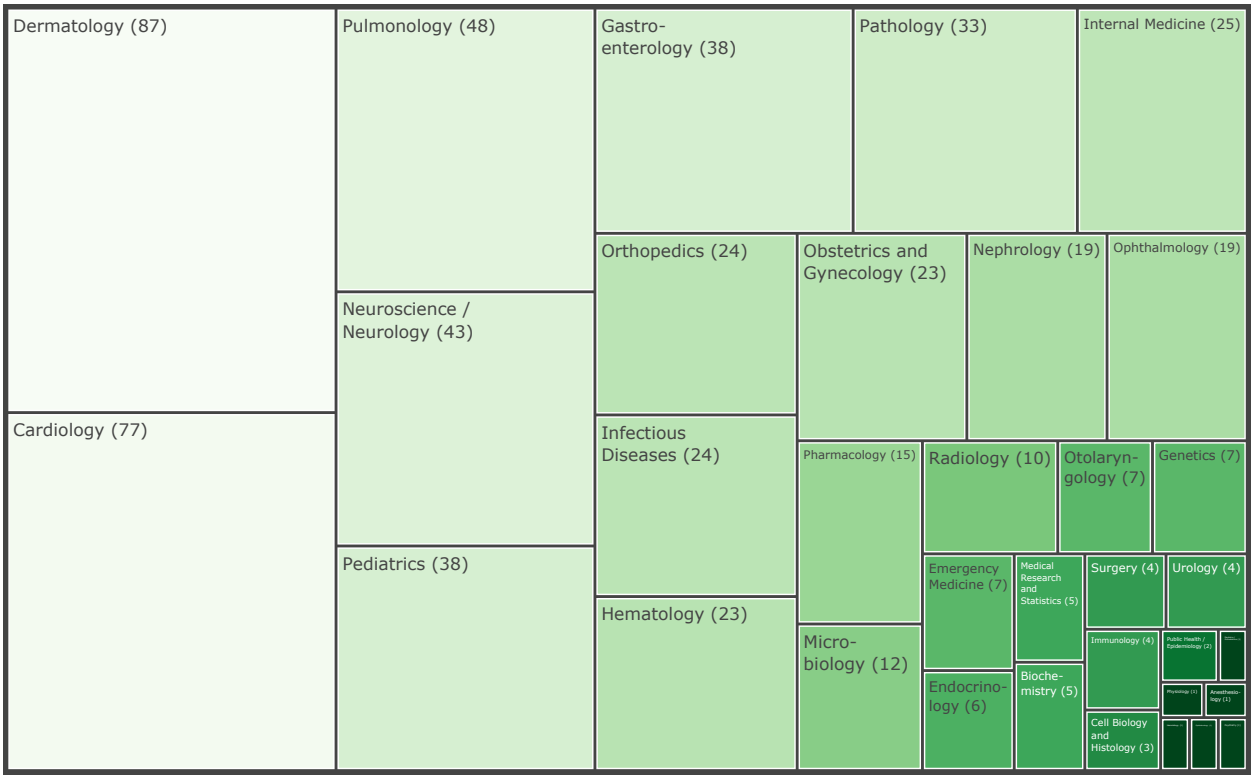

Figure 7: Distribution of specialty topics in the Visual USMLE dataset, as classified by Claude-1 using the categories provided in Table 2.

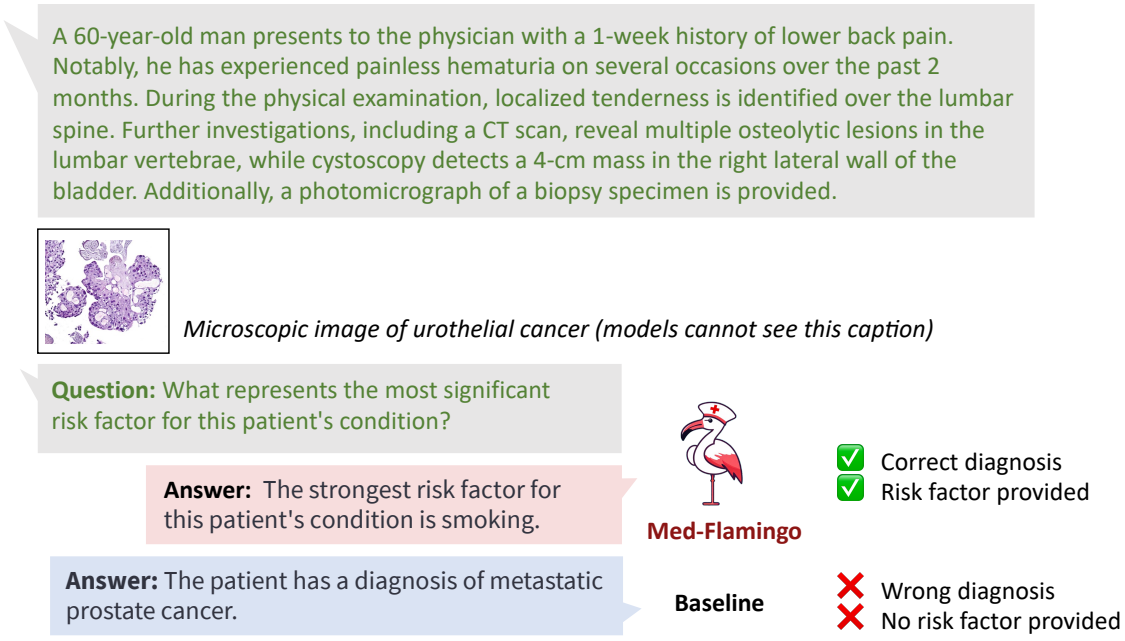

Figure 8: Example of a Visual USMLE problem. The displayed baseline answer is from the OpenFlamingo model.