

# Med-Flamingo: a MultiModal Medical Few-Shot Learner.

## Key Ideas

Flamingo is a VLM that is trained on multi-modal interleaved text-image data that generates text as an output. The paper leverages Flamingo, a Visual Language Model for medical applications.

## Flamingo

Flamingo paper talks about the shortcomings of some of the multi-modal few-shot architectures (Late 2022) with respect to their performance issues and introduce Flamingo, a VLM that is a very good **few-shot learner**. It is based on the idea of conditioning a very good language model with visual features.

### Architecture:

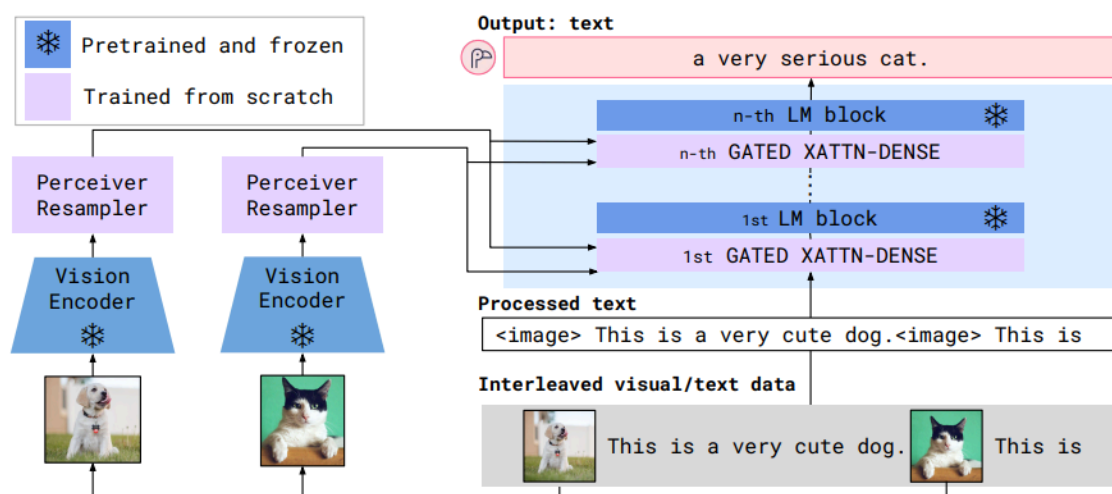


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

The Flamingo architecture has 4 important parts:

- **Vision encoder:** This could be any vision based encoder to extract visual features. The vision encoder is always frozen.
- **LLM:** A frozen language model for generating open-ended text based output
- **Perceiver resampler:** A perceiver based resampler, that makes sure that multi-modal input always produces the fixed number of output tokens. It uses latent queries to extract information from varied size visual keys and values. (For image = video with 1 frame; the time and space dimensions are collapsed into one before sending it to the perceiver)
- **Gated Cross-attention :** A cross attention between the visual and the text tokens, gated by a tan-h function.

#### **It works as follows:**

Image/video is encoded by the image encoder, post which it the tokens are sent into the perceiver sampler. In the perceiver sampler, which is based on perceiver architecture, will take in variable number of image/video tokens and returns a fixed output of 64 tokens.

Text is tokenized using text tokenizers. A masked gated cross-attention is applied between vision tokens and text tokens. The mask is applied in such a way that the text tokens only attend to the image preceding it.

These attention blocks are interleaved with the LLM blocks

#### ▼ Gated Masked Cross-Attention

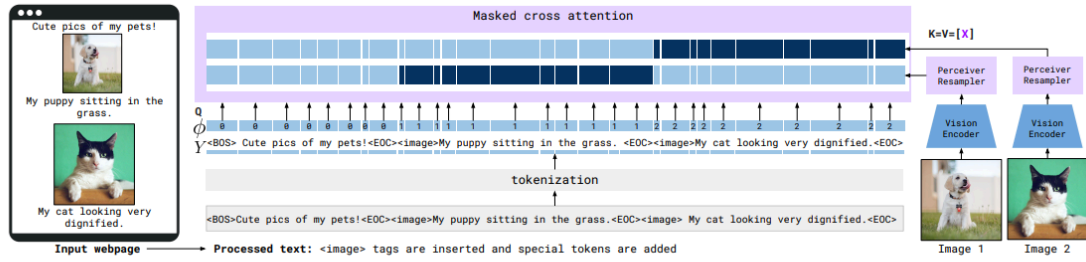


Figure 7: **Interleaved visual data and text support.** Given text interleaved with images/videos, e.g. coming from a webpage, we first process the text by inserting `<image>` tags at the locations of the visual data in the text as well as special tokens (`<BOS>` for “beginning of sequence” or `<EOC>` for “end of chunk”). Images are processed independently by the Vision Encoder and Perceiver Resampler to extract visual tokens. At a given text token, the model only cross-attends to the visual tokens corresponding to the last preceding image/video.  $\phi$  indicates which image/video a text token can attend or 0 when no image/video is preceding. In practice, this selective cross-attention is achieved through masking – illustrated here with the dark blue entries (unmasked/visible) and light blue entries (masked).

Flamingo is trained on datasets with image-text pairs, video-text pairs and multiple images-text pairs.

#### ▼ Training datasets structure

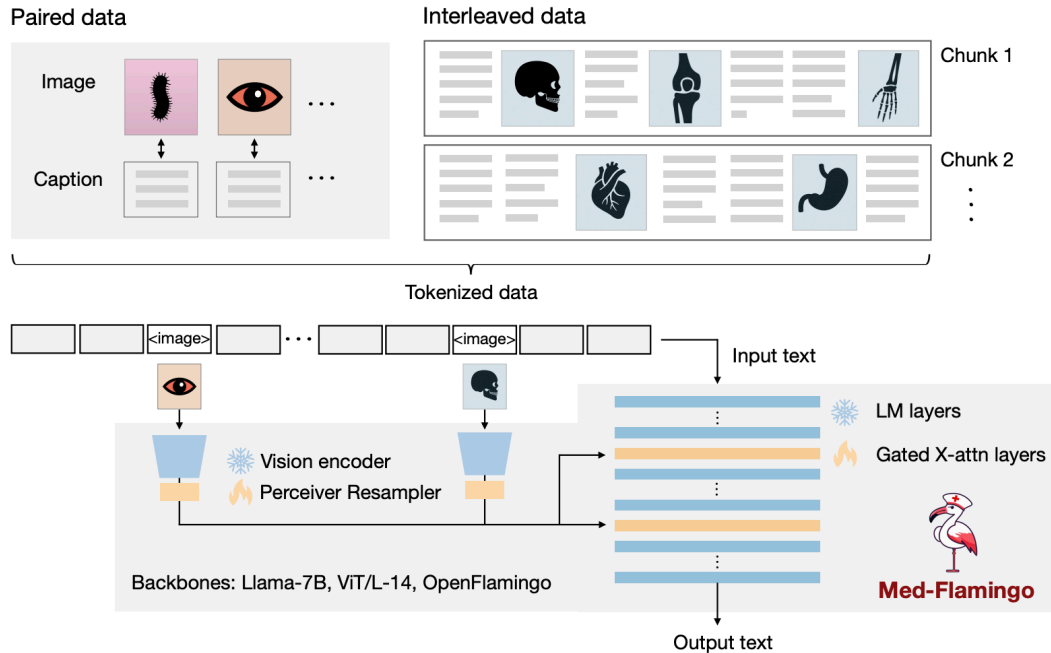


Figure 9: **Training datasets.** Mixture of training datasets of different formats.  $N$  corresponds to the number of visual inputs for a single example. For paired image (or video) and text datasets,  $N = 1$ .  $T$  is the number of video frames ( $T = 1$  for images).  $H$ ,  $W$ , and  $C$  are height, width and color channels.

while Flamingo is good on general tasks and show good in-context learning (few shot) performance on some specified tasks, it lacked medical knowledge since it was not trained directly on any medical datasets.

Med-Flamingo is trained on vast set of medical dataset. The dataset is of 2 types. Image-text pairs and image-text interleaved data

## 1. Multimodal pre-training on medical literature



### Training Datasets:

**MTB** : A multimodal dataset created from ~4721 medical textbooks from different medical specialities. 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

**PMC-OA**: A bio-medical data with image-caption pairs collected from PubMedCentral's OpenAccess subset.

## Impressions

The med-VQA s evaluated on 2 datasets : **VQA-RAD** , **PathVQA**

Furthermore, they also created a visual-USMLE (united nations medical licensing exam) and changed the questions from multiple choice to open-ended

### Evaluation Metrics:

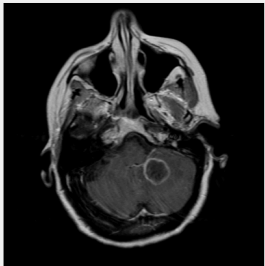
The VAQ task was tested using the following metrics:

1. **Clinical evaluation score** : Clinicians take a look at the predictions of the model for a given task and rate the prediction out of 10.

▼ Example

**Human Evaluation App**
Problem 1/50

Image:



Question: Where is the lesion located?

Correct answer: left cerebellum

prediction\_1: left cerebellum

Quality score from 0 to 10

10

0

10

0

prediction\_2: in the patient's left temporal lobe

10

0

10

0

prediction\_3: in the left cerebellum

10

0

10

0

prediction\_4: The left frontal lobe

10

0

10

0

Next Problem

2. **Bert Similarity Score**

3. **Exact-match** : Generated text should exactly match the ground truth.

\*\*MedVINT is a VQA that is used as a baseline. It has no few shot capability. Hence it is finetuned on the training splits of the VQA datasets.

**Performance on VQA-RAD:**

| VQA-RAD                                 | Clinical eval. score | BERT-sim     | Exact-match  |
|---|----------------------|--------------|--------------|
| 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        | <b>0.200</b> |
| Med-Flamingo zero-shot                  | 3.82                 | 0.480        | 0.000        |
| Med-Flamingo few-shot                   | <b>5.61</b>          | <b>0.650</b> | <b>0.200</b> |

Med-Flamingo shows good performance with few shot examples across. But we also notice that openFlamingo also performs nearly similar to Med-

Flamingo even having been not trained on the medical datasets explicitly. This shows the in-context learning capabilities of the Flamingo models.

### Performance on Path-VQA:

| Path-VQA                                 | Clinical eval. score | BERT-sim     | Exact-match  |
|--|----------------------|--------------|--------------|
| MedVINT zero-shot                        | 0.13                 | 0.608        | 0.272        |
| MedVINT fine-tuned ( $\sim 20K$ samples) | 1.23                 | <b>0.723</b> | <b>0.385</b> |
| OpenFlamingo zero-shot                   | <b>2.16</b>          | 0.474        | 0.009        |
| OpenFlamingo few-shot                    | <u>2.08</u>          | 0.669        | 0.288        |
| Med-Flamingo zero-shot                   | 1.72                 | 0.521        | 0.120        |
| Med-Flamingo few-shot                    | 1.81                 | <u>0.678</u> | <u>0.303</u> |

Its the same story as above here. We see openFlamingo performing better than med-Flamingo

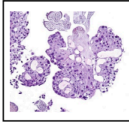
### Visual USMLE:

| Visual USMLE           | Clinical eval. score | BERT-sim     |
|------------------------|----------------------|--------------|
| MedVINT zero-shot      | 0.41                 | 0.421        |
| OpenFlamingo zero-shot | <u>4.31</u>          | <b>0.512</b> |
| OpenFlamingo few-shot  | 3.39                 | 0.470        |
| Med-Flamingo zero-shot | 4.18                 | <u>0.473</u> |
| Med-Flamingo few-shot  | <b>4.33</b>          | 0.431        |

### Qualitative Examples:

Example: Doesn't give out diagnosis and answers the asked question correctly

A 60-year-old man presents to the physician with a 1-week history of lower back pain. Notably, he has experienced painless hematuria on several occasions over the past 2 months. During the physical examination, localized tenderness is identified over the lumbar spine. Further investigations, including a CT scan, reveal multiple osteolytic lesions in the lumbar vertebrae, while cystoscopy detects a 4-cm mass in the right lateral wall of the bladder. Additionally, a photomicrograph of a biopsy specimen is provided.



*Microscopic image of urothelial cancer (models cannot see this caption)*

**Question:** What represents the most significant risk factor for this patient's condition?

**Answer:** The strongest risk factor for this patient's condition is smoking.

**Answer:** The patient has a diagnosis of metastatic prostate cancer.



**Med-Flamingo**

✓ Correct diagnosis  
✓ Risk factor provided

**Baseline**

✗ Wrong diagnosis  
✗ No risk factor provided

Figure 6: Example of a Visual USMLE problem.

Example: Few shot

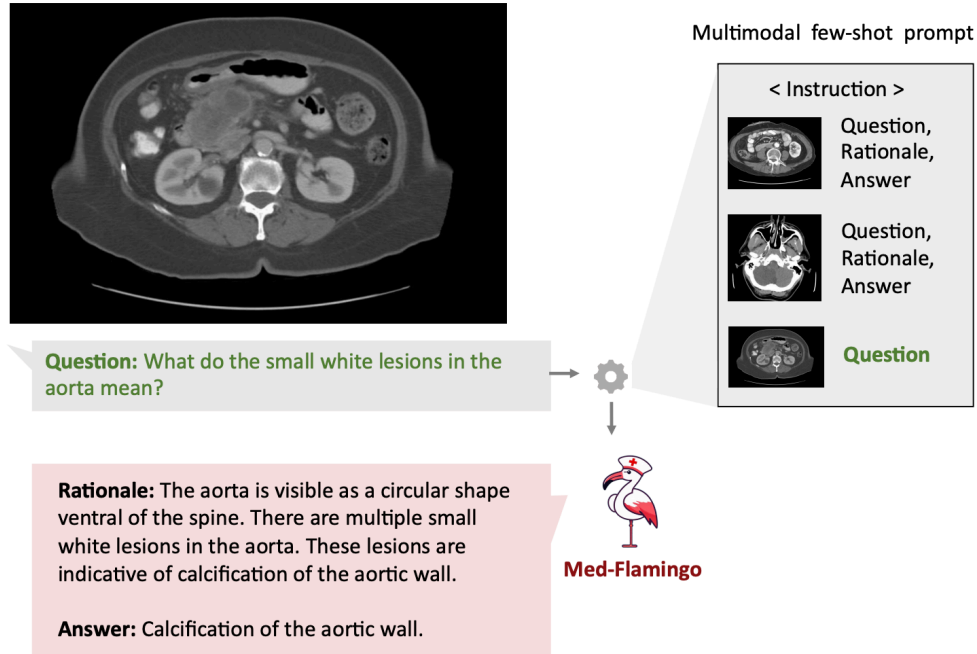


Figure 5: Multimodal medical few-shot prompting illustrated with an example. Few-shot prompting here allows users to customize the response format, *e.g.*, to provide rationales for the provided answers. In addition, multimodal few-shot prompts potentially offer the ability to include relevant context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

## Example: Correct diagnosis

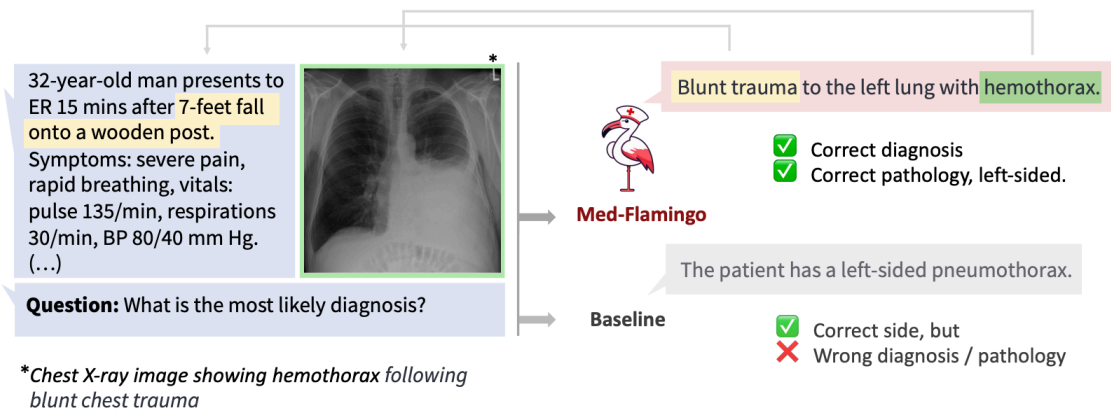


Figure 1: Example of how Med-Flamingo answers complex multimodal medical questions by generating open-ended responses conditioned on textual and visual information.